DOCUMENT RESUME

ED 171 732	TM 008 684

AUTHOR	Harris, Chester W., Ed.; And Others
TITLE	Problems in Criterion-Referenced Measurement. CSE Monograph Series in Evaluation, 3.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
PUB DATE	74
NOTE	180p.
AVAILABLE FROM	Center for the Sudy of Evaluation, University of California, Los Angeles, California 90024 ($3.50)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	*Achievement Tests; Cognitive Objectives; *Criterion Referenced Tests; *Educational Objectives; Evaluation Criteria; Mastery Tests; Measurement Goals; Psychometrics; *Test Construction; *Testing Problems; Test Validity

ABSTRACT

Six essays on technical measurement problems in criterion referenced tests and four essays by psychometricians proposing solutions are presented: (1) "Criterion-Referenced Measurement" and Other Such Terms, by Marvin C. Alkin which is an overview of the first six papers; (2) Selecting Objectives and Generating Test Items for Objectives-Based Tests, by W. James Popham; (3) A Judgmental Approach to Criterion-Referenced Testing, by H. A. Wilson; (4) Measurement Considerations in Instructional Product Development, by Robert L. Baker; (5) Generating Criterion-Referenced Tests from Objectives-Based Assessment Systems: Unsolved Problems in Test Development, Assembly, and Interpretation, by Rodney W. Skager; (6) Problems in the Development of Criterion-Referenced Tests: The IPI Pittsburgh Experience, by Anthony J. Nitko; (7) Problems of Objectives-Based Measurement, by Chester W. Harris; (8) Thus Spake Psychometrika, by W. James Popham, an overview of the psychometricians' contributions; (9) Some Technical Characteristics of Mastery Tests, by Chester W. Harris; (10) The Preparation of Criterion-Referenced Tests, by Frederick B. Davis and James J. Diamond; (11) Prescribing Test Length for Criterion-Referenced Measurement, by Melvin R. Novick and Charles Lewis; (12) Empirical Validation of Criterion-Referenced Measures, by J. Ward Kessling. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED171732

T M
Reg

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CSE MONOGRAPH SERIES IN EVALUATION

3

PROBLEMS IN CRITERION-REFERENCED MEASUREMENT

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Mark Young
CSE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

CENTER FOR THE STUDY OF EVALUATION
UNIVERSITY OF CALIFORNIA • LOS ANGELES

TM008 684

CSE MONOGRAPH SERIES IN EVALUATION

SERIES EDITOR
Marvin C. Alkin

Center for the Study of Evaluation
UCLA Graduate School of Education
University of California, Los Angeles
Los Angeles, California 90024

PROBLEMS IN CRITERION-REFERENCED MEASUREMENT

Edited by

Chester W. Harris
University of California, Santa Barbara

Marvin C. Alkin
Center for the Study of Evaluation

W. James Popham
University of California, Los Angeles

Center for the Study of Evaluation
University of California, Los Angeles, 1974

4

CSE MONOGRAPH SERIES IN EVALUATION

NUMBER

1. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the MINNEMAST Project
   Wells Hively, Graham Maxwell, George Rabehl, Donald Sension, and Stephen Lundin $3.50
2. National Priorities for Elementary Education
   Ralph Hoepfner, Paul A. Bradley, and William J. Doherty $3.50
3. Problems in Criterion-Referenced Measurement
   Chester W. Harris, Marvin C. Alkin, and W. James Popham (Editors) $3.50

TABLE OF CONTENTS

Preface vii

SECTION I

"Criterion-Referenced Measurement" and Other Such Terms 3
Marvin C. Alkin

Selecting Objectives and Generating Test Items
for Objectives-Based Tests 13
W. James Popham

A Judgmental Approach to Criterion-Referenced Testing 26
H.A. Wilson

Measurement Considerations in Instructional Product Development 37
Robert L. Baker

Generating Criterion-Referenced Tests from Objectives-Based Assessment
Systems: Unsolved Problems in Test Development, Assembly, and
Interpretation 47
Rodney W. Skager

Problems in the Development of Criterion-Referenced Tests: The IPI
Pittsburgh Experience 59
Anthony J. Nitko

Problems of Objectives-Based Measurement 83
Chester W. Harris

SECTION II

Thus Spake Psychometrika . . . 95
W. James Popham

Some Technical Characteristics of Mastery Tests 98
Chester W. Harris

The Preparation of Criterion-Referenced Tests 116
Frederick B. Davis and James J. Diamond

Prescribing Test Length for Criterion-Referenced Measurement 139
Melvin R. Novick and Charles Lewis

Empirical Validation of Criterion-Referenced Measures 159
J. Ward Keesling

6

PREFACE

While the clamor from American educators for criterion-referenced tests has been discernibly increasing in recent years, this increase has been matched by the needs for technical procedures to produce such criterion-referenced measures. In an effort to foster a more enlightened dialogue on these technical problems, the project leading to the series of essays in this volume was initiated early in 1972. A brief account of its major elements may assist the reader in understanding the format of the volume and the sequential manner in which its components were prepared.

The design for the monograph was originally conceived as the result of a planning session involving the monograph's three editors and Richard E. Schutz, Director of the Southwest Regional Laboratory for Educational Research and Development. Because we recognized that a number of criterion-referenced measurement projects were underway in which technical advances would be welcomed, we decided to try to identify the nature of the technological problem areas by asking a group of practitioners to set forth their pressing technical measurement problems related to criterion-referenced testing.

These papers were prepared, and the senior editor was asked to survey the concerns they represented and then select a group of psychometricians who would propose solutions to the problems perceived by the practitioners. The monograph would, then, consist of two different sections prepared by two groups of individuals with essentially different orientations. Although the psychometricians obviously could not deal with every concern registered by the practitioners, it was hoped that they would focus their remarks in terms of the practitioners' problems.

Because their task was considered to be more difficult, the psychometricians were provided with a modest honorarium for preparing their papers. Division D of the American Educational Research Association, consistent with its measurement and research methodology focus, provided the funds for two of the four psychometricians' honoraria. While AERA will share in any royalties resulting from this publication, the monograph was not reviewed by the AERA Publications Committee, hence no official AERA endorsement of the volume should be inferred.

With these remarks as a backdrop for the volume, the editors hope the reader will find the following essays consistent with our wish to sharpen the discussion of criterion-referenced technology.

C.W.H.
M.C.A.
W.J.P.

March, 1974

SECTION I

"CRITERION-REFERENCED MEASUREMENT" AND OTHER SUCH TERMS

Marvin C. Alkin
Center for the Study of Evaluation

Criterion-referenced measurement is a concept which has tantalized and frustrated many educators. Recent problems encountered in the development and utilization of criterion-referenced measurement systems have demonstrated that the problem cannot be viewed simply in terms of user inability to comprehend what is available; there is also a host of complex technical problems associated with the development itself. In the first section of this monograph, a number of leading educational professionals engaged in the development of criterion-referenced measurement systems describe some current developmental activities so as to illustrate the methodological obstacles met in the construction and use of criterion-referenced tests. Some of the problems identified are then used as foci of attention in the papers comprising the second part of the monograph.

In a new field where many researchers are working concurrently and more or less independently, one can expect to find a great deal of new terminology reflecting different perspectives of the conceptual landscape. Such is the case with criterion-referenced measurement. To introduce the first section of the monograph and the field of criterion-referenced testing in general, this overview relates these various terms to one another and identifies their origins in educational and psychometric literature.

This introduction addresses four major topics of concern in the field of criterion-referenced assessment, namely: (1) *the term "criterion-referenced tests"*; (2) *domain specification and item forms;* (3) *specifying and organizing instructional goals;* and (4) *mastery testing.*

CRITERION-REFERENCED TESTS

Great attention has recently been directed to the types of tests appropriate for different assessment and information purposes. One frequent use of a test is to assess whether an individual examinee has achieved some prescribed degree of competence on an instructional task, usually referenced to a performance objective or a behavioral objective. This test information can then be used for assessment purposes to evaluate the student's mastery of behavioral objectives characteristically associated with a specific curriculum or textbook, or for diagnostic purposes to place a

---

Preparation of this paper was supported by funds from the National Institute of Education (NIE), Department of Health, Education, and Welfare. Points of view or opinions expressed do not necessarily represent official NIE position or policy.

student in an appropriate instructional unit. Tests designed for this purpose are known as objectives-based or criterion-referenced tests.

A frequent question about criterion-referenced tests refers to their relationship to the classical "norm-referenced" test. In brief, the crucial difference between these tests is that criterion-referenced tests, unlike norm-referenced tests, are not intended to compare or rank individuals and are developed from well-defined performance domains or objectives. Popham and Husek (1969) and Hambleton and Novick (1972) have provided in-depth discussions of the measurement implications of these differences.

The term "criterion-referenced test" (CRT) has been defined in a variety of ways in the literature. Three of the most widely applied definitions are listed below:

1. *A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards* . . . Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. (Glaser & Nitko, 1971, p. 653)
2. A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed. (Harris & Stewart, 1971)
3. Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., a performance standard. (Popham & Husek, 1969, p. 2)

While these definitions differ considerably in terms of the limitations and constraints placed on a criterion-referenced measure, they all share a common emphasis on two characteristics. First, *each* definition *emphasizes test organization* (i.e., test-item selection) based on specific tasks or behavioral objectives. Second, *each* definition *emphasizes assessment in terms of predefined* performance criteria.

These two characteristics are essential to each of the CRT systems discussed in the first section of this monograph. However, there is some variation in the emphasis placed on each of the characteristics. Nitko, for example, focuses greater attention on the second characteristic: "CRT, then, is one that is deliberately constructed to give scores that tell what kinds of behavior individuals with those scores can demonstrate." Skager emphasizes the first characteristic (test organization): ". . . criterion-referenced tests, whatever their other characteristics, are by definition 'objectives-based'." Thus, although it may be noted that the terms "criterion-referenced" and "objectives-based" are typically used interchangeably, in some cases "criterion-referenced" is used as a more general term and refers to a test that is associated with instructional performance criteria, while "objectives-based" refers to tests referenced especially to behavioral objectives defined in great detail.

The following is a listing of relevant terms used by each of the authors writing in this section of the monograph:

Popham:	objectives-based tests
Wilson:	criterion-referenced testing
Baker:	criterion-referenced tests
Skager:	criterion-referenced tests objectives-based tests objectives-based assessment systems
Nitko:	criterion-referenced tests criterion-referenced score interpretations

DOMAIN SPECIFICATION AND ITEM FORMS

Inherent in the concept of a criterion-referenced measurement system are performance criteria. Consequently, one important problem for the test developer is the specification and organization of relevant behaviors that define these performance criteria. Several systems have been suggested or are currently being developed to meet this need, resulting in a proliferation of terminology (as well as methodologies).

The basis for many such systems used for the specification and organization of relevant behaviors is the work of Wells Hively and his colleagues at the MINNEMAST Project. Hively's system, frequently referred to as "domain-referenced" measurement or theory of performance, emphasizes careful definition of the domain of relevant learner behaviors associated with an area of knowledge, and subsequent referencing of test items to this domain.

Hively sought to bridge the gap that exists between the statement of a behavioral objective (the criterion) and the criterion-referenced tests constructed to measure the achievement of such criteria by specifying the domain of behaviors relevant to specific areas of knowledge and then developing rules for generating sets of equivalent items that represent these skills and concepts. In Hively's words:

> The basic notion underlying domain-referenced achievement testing is that certain important classes of behavior in the repertoires of experts (or amateurs) can be exhaustively defined in terms of structured sets or *domains* of test items. Testing systems may be *referenced* to these domains in the sense that a testing system consists of rules for sampling items from a domain and administering them to an individual (or sample of individuals from a specified population) in order to obtain estimates of the probability that the individual (or group of individuals) could answer any given item from the domain at a specified moment in time.
>
> Domains of test items are structured and built up through the specifi-

> cation of stimulus and response properties which are thought to be important in shaping the behavior of individuals who are in the process of learning to be experts. These properties may be thought of as stratifying large domains into smaller domains or subsets.
>
> Precise definition of a domain and its subsets makes statistical estimation possible. This provides the foundation for precise diagnosis of the performance of individuals over the domain and its subsets. In addition, clear specification of the properties used to structure the domain makes possible inductive generalization beyond the domain to situations which share those properties. That is, once we have diagnosed a student with respect to a defined domain we may be able to predict his behavior (in a non-statistical, inductive fashion) in natural situations which have some properties in common with the test items within the domain. (Hively, et al., 1973, p. 15)

Rules for generating items representing a given learning domain are organized into formal schemes called "item forms" that specify the elements necessary to produce the items needed to assess some objective(s). There are three major components to an item form: (1) instructions, (2) stimulus characteristics, and (3) response characteristics. Instructions delimit the specific directions given to students and provide any needed information not supplied by the actual items. Stimulus characteristics define rules for constructing items. The parts of an item that must remain invariant as well as those that can vary are specified and the eligible replacement possibilities are indicated. The common unvarying components of all items generated by an item form are contained in the item form shell and possible replacements for these invariant components are identified by replacement sets. Finally, the acceptable ways of responding to items (e. g., written, oral, multiple choice, essay . . .) and definitions of correct responses are supplied by the response characteristics.

It should be noted that Hively's concept of a domain includes specific content areas *as well as* behaviors associated with this content. There are two other prominent conceptions of a domain that should be noted. Cronbach's (1971) conception of "universe specification" focuses on skills and Ebel's (1962) conception of a "standard domain of content" focuses on content. Cronbach suggests that the universe specification presumably will define a category of skills to be tallied and a list of situations in which observations are to be made. Each observation, then, should provide a valid sample of this universe. Ebel developed the concept of a standard domain of content to provide a foundation for test construction. Tests developed from a standard domain of content would be based directly on the tasks which make up or define some prespecified content area. Such tests would be standardized in the sense that the standard domain of content would provide a basis for creating equivalent items.

Domain-referenced testing and its associated item-form analysis (Hively, et al., 1973) serve as a focus for much of the work and terminology represented in this monograph.[1] Baker uses the terms "behavioral classes" and "behavioral content domains" in a sense very similar to Hively's domain specification. In addition, Baker uses the item form concept. Nitko also considers domain specification using the closely related term "behavior classes" that define various levels of competence. He describes techniques for specifying these domains, called "domain descriptors," in a manner which follows from Hively's work. Popham, who speaks of "content-general domains," has developed a more simplified system of generating items based on item-form analysis. Finally, Skager uses the terms "performance domains" and "content domains" in a "Hively" sense.

The universe concept of Cronbach is referenced by Nitko while Baker uses the parallel term "universe of content." Wilson speaks of a "domain of reference" that defines a subject matter area and the associated "universe of behaviors" that identifies the important associated learner behaviors.

To summarize, in terms of the concept of domain specification, the following terms are used by the authors of this section:

Popham:	domain-referenced testing (Hively) item forms content general domains
Wilson:	domain of influence universe of behaviors
Baker:	behavioral classes item forms (Hively) universe of content behavioral-content domains
Skager:	performance domains content domain
Nitko:	standard domain of content theory of performance (Hively) item form analysis domain specification domain descriptors classes of behavior universe specification

SPECIFYING AND ORGANIZING INSTRUCTIONAL GOALS

During the 1960's the education community accepted and followed the principle of presenting its instructional goals in precise behavioral terms.

---

[1] A full explication of domain-referenced testing is presented in W. Hively, G. Maxwell, G. Rabehl, D. Sension, & S. Lundin, *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project.* CSE Monograph Series in Evaluation, No. 1 (Los Angeles: Center for the Study of Evaluation, University of California, 1973).

It was argued that in order to facilitate evaluation and to account for the outcomes of education, objectives must be stated in terms of measurable learner behaviors. These "behavioral objectives" were intended to reduce the ambiguity of educational intentions by specifying the activities and conditions which attend each desired outcome.

Just as the use of objectives stated in behavioral terms is widespread, so the *names* used to refer to such objectives are equally abundant. Some of the more prominent names include behavioral objectives, performance objectives, prespecified instructional outcomes, and behavioral referents. Behavioral output (Skager), outcomes orientation, and consequence orientation (Popham) are also terms closely related to behavioral objectives which focus on the outcomes of instruction.

One of the most perplexing and as yet unresolved issues in stating instructional objectives is the optimum degree of generality to be employed. Objectives can be stated so precisely that they merely parallel test items or, on the other hand, can be stated so generally that they lead to different interpretations and are thereby rendered useless for constructing equivalent tests. The breadth or narrowness of an objective's coverage is partly determined by the elements it includes. A definition of *conditions* under which measurement of the objective takes place (e.g., open vs. closed book; in front of a student audience vs. into a tape recorder) and the specification of *standards* of performance to be reached in order for the objective to be achieved (e.g., 80% correct; in less than two minutes) are two such elements which may be included (Mager, 1962; Popham, 1965). Another aspect of specificity concerns the characteristics of the area to be assessed and, in particular, the place of a given objective in some logical or sequential organization of related objectives. The setting for and organization of a collection of objectives may largely determine how they should be stated. In order to cope with the problem of what an objective should contain, various systems of defining the generality of objectives have been proposed. Many of these proposed systems (two of which are discussed in this monograph) establish a hierarchy delimiting various levels of specificity or generality.

Popham, one of the original proponents of behaviorally-stated objectives, suggests a four-level hierarchy which is used at the Instructional Objectives Exchange (IOX). The most general objectives, called "major categories," usually describe important and often comprehensive skills such as using sets (in mathematics) or verb tenses (in grammar). Content general objectives then describe intermediate skills within each category, for example, intersection-union principles of set theory or past participle usage. The precise statement of a skill is the specific "objective" itself. Finally, the construction of test items is based on "amplified objectives" which are defined as expanded objectives containing sufficient detail regarding the nature of measurement procedures to facilitate item development.

Wilson introduces a three-level generality hierarchy which is used at the National Assessment of Educational Progress. Educational goals determined through a national goal consensus are stated as "overall objectives," for example, "be able to perform basic arithmetic computations." Specific content areas and behaviors are then further delimited by major objectives, for example, "be able to divide and multiply decimal and integer numbers." Finally, various levels of "sub-objectives" define precise performance criteria which can be used for item development, for example, "be able to divide a 7-digit (or less) number with up to three decimal values by a 10-digit (or less) integer."

The following are the terms used with reference to instructional goals by the authors of this section of the monograph:

Popham:	instructional objectives outcomes orientation consequence orientation major categories content general objectives objectives amplified objectives
Wilson:	overall objectives major objectives sub-objectives
Baker:	instructional objectives prespecified instructional outcomes
Skager:	performance objectives behavioral output
Nitko:	behavioral objectives

MASTERY TESTING

One of the essential features of criterion-referenced testing is its emphasis on assessment in terms of prespecified behavioral objectives, thereby providing a means for describing what students can do or what they know or what they think without reference to the skills, knowledge, or attitudes of others (Klein & Kosecoff, 1973). This distinctive feature of criterion-referenced tests leads to the reporting of test scores in terms of absolute (rather than relative) measures. The test scores define a level of performance or mastery of an objective or skill. Some of the different ways of reporting criterion-referenced test results are discussed by Harris, who identifies five "directly interpretable" measurement scales. By "directly interpretable," Harris means a measurement scale that has precise meaning without reference to the scores of other individuals (without reference to norms). The directly interpretable measurement scales (or "metrics") are:

1. *rate metric*—the time it takes for a student to complete a given task.
2. *sign metric*—the accomplishment or not ("on-off" in nature) of a task. A student is either a master or not (as for a single-item test).
3. *accuracy metric*—proportion of times that a student is successful.
4. *proportion metric*—the portion of a population of items from a well-defined domain, which a student can perform or knows.
5. *scaling metric*—an estimation of a student's performance level derived by scaling item responses in a Guttman-like sense.

Of all the interpretations listed above, the one that has been the focus of most discussion is that of mastery (the sign metric). One of the reasons for this attention is that the concept of mastery comes closest to the underlying spirit of criterion-referenced testing and is most appropriate for measures of cognitive and attitudinal skills.

The notion of mastery is certainly very appealing; learning is a purposeful activity and its results need not be ranked according to a normal curve. It would be preferable, rather, to set educationally significant standards and attempt to have a majority of the student population meet these goals. Several theories of mastery learning have been proposed, for example, Bruner (1966), Carroll (1963), Goodlad and Anderson (1959). Bloom (1968) summarized the basic premise uniting these approaches:

> . . . if students are normally distributed with respect to aptitude, but the kind and quality of instruction and the amount of time available for learning are made appropriate to the characteristics and needs of *each* student, the majority of students may be expected to achieve mastery of the subject (p. 3).

Despite the great appeal of the mastery concept, however, several problems remain unresolved. Most important, present psychometrics and theories of mastery learning have not provided a means of establishing an educationally useful definition of mastery. Indeed, many proponents of mastery learning and mastery testing, while carefully attending to the details of implementing their systems and their potential impact on teachers and students, have given little consideration to this important area. Arbitrary performance standards such as 85% correct responses are common, but rarely is there any satisfactory criterion for establishing a mastery standard. Even Bloom, who ardently supports teaching to mastery, conceded that "While we would recommend the use of absolute standards carefully worked out for a subject, we recognize the difficulty of arriving at such standards (Bloom, 1968, p. 8)" and in the absence of absolute criteria recommends standards based on previous experience.

It is further argued (e. g., Cronbach, 1971) that a single mastery score may be inadequate; that absolute scores are not appropriate for the wide range of student aptitudes and needs. Further, a single mastery score can hide the true level of achievement (in other words, if the student failed to meet mastery criteria, did he miss by a great deal and, if so, how does he differ from a student who just squeaks by?).

Despite these difficulties, the notion of mastery remains an integral part of many criterion-referenced testing systems. Some consideration of mastery is reflected in several of the papers in this section of the monograph. Skager voices many of the aforementioned misgivings when he considers how many items are needed to measure the "predetermined, but theoretically unsubstantiated criterion of mastery." Nitko asserts that the concept of a "desired model or minimum goal" is basic to the nature of a criterion-referenced test. To that end, he offers a definition of mastery in his paper. Finally, Baker discusses "specified instructional decision algorithms" (that is, rules for determining mastery). The importance of this concept is further emphasized in Baker's use of the term "mastery items" for items providing "practice and assessment of desired behavior" and a "Learning Mastery System" for SWRL's instructional management system.

The terms used by the various authors dealing with the topic of mastery are as follows:

Baker:	specified instructional decision algorithms mastery items Learning Mastery System
Skager:	mastery, criterion of mastery
Nitko:	desired model or minimum goal mastery mastery testing

Although the terms used throughout these articles are often unique to an individual author, it is clear that similar issues are addressed in all of the papers even while individual papers reflect different perspectives toward these issues. It is hoped that the articles in this section of the monograph will stimulate further attempts to resolve the methodological problems posed.

REFERENCES

Bloom, B.S. Learning for mastery. *Evaluation Comment*, 1968, *1*(2).

Bruner, J.S. *Toward a theory of instruction*. Cambridge, Mass.: Harvard University Press, 1966.

Carroll, J. A model of school learning. *Teachers College Record*, 1963, *64*, 723–733.

Cronbach, L.J. Test validation. In R. L. Thorndike (Ed.), *Educational*

Measurement. (2nd ed.) Washington: American Council on Education, 1971.

Ebel, R. Content-standard test scores. *Educational and Psychological Measurement*, 1962, 22, 15–25.

Eisner, E.W. Educational objectives: Hindrance? *School Review*, 1967, 75, 250–266.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational Measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Goodlad, J.I., & Anderson, R.H. *The nongraded elementary school*. New York: Harcourt, Brace, 1959.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. *ACT Research Report* No. 53. Iowa City, Iowa: The American College Testing Program, 1972.

Harris, M.L., & Stewart, D.M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hively, W. Domain-referenced achievement testing. A paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project*. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Klein, S.P., & Kosecoff, J.B. Issues and procedures in the development of criterion-referenced tests. *ERIC/TM Report* 26. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement and Evaluation, 1973.

Mager, R.F. *Preparing instructional objectives*. San Francisco: Fearon, 1962.

Popham, W.J. *The teacher-empiricist; A curriculum and instruction supplement*. Los Angeles: Lennox-Brown, 1965.

Popham, W.J., & Baker, E.L. *Establishing instructional goals*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Popham, W.J., & Husek, T.R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1–9.

SELECTING OBJECTIVES AND GENERATING TEST ITEMS FOR OBJECTIVES-BASED TESTS

W. James Popham
University of California, Los Angeles

This paper deals with considerations stemming from a decision to prepare sets of objectives-based tests to be distributed by the Instructional Objectives Exchange (IOX) commencing in February, 1973. One or two brief statements regarding the IOX decision to develop objectives-based tests may prove helpful. Since 1968 the Exchange has been developing and distributing booklets containing collections of measurable instructional objectives and related test items. The chief purpose of this activity has been to encourage American educators to organize their instructional activities around explicitly stated instructional goals, thereby attending to the *consequences* of instructional procedures rather than to the procedures themselves. By providing objectives collections from which educators could select, we hoped to make their task easier, and thus incline them to employ an outcomes orientation in their work.

During the early years of IOX operations it became apparent that some educators did indeed use the collections as we hoped; that is, by selecting objectives and constructing tests largely on the basis of the sample items we supplied. Yet many people still felt that this was too much work, and probably did little more with the collections than place them in the school district's curriculum library. We had to make it easier still for such educators to adopt a consequence orientation. The decision to develop objectives-based tests represented an effort to provide readily used support materials whereby a teacher would discover how much pupils had learned. If teachers had access to brief, easily used tests which could be employed both to diagnose learner deficiencies and reflect mastery of important objectives, we believed that those teachers would really use such tests.

We wished to avoid one of the chief deficiencies of standardized tests, namely, measuring a general construct with such vagueness that even when a teacher discovers that pupils are weak in a given area, the nature of the deficit is so indeterminate that little can be done to ameliorate it. We decided to base the items in our tests, therefore, on extremely explicit statements of instructional intent referred to as *amplified objectives*.

In general, our tests will be distributed on *one-page* pre-printed spirit masters which will be used by teachers to duplicate sufficient copies of tests for their students. While it is possible to have two or more pages per test, the typical test will be a one-page sheet with five or ten items per test, although in rare instances we might have two easily measured objectives on one page.

Cost considerations have dictated that this spirit master distribution scheme will be our typical format. Decisions regarding objectives and test items will have to be made in relationship to this constraint.

A series of preliminary decisions regarding the form of objectives-based tests has led to the necessity of generating a partial set of guidelines for use by those individuals who will devise these tests. This paper deals with these guidelines and concludes with a set of unresolved problems associated with this sort of measurement construction enterprise.

Prominent among the problems emerging from these decisions are the questions of: (1) Which objectives should be selected for inclusion in the tests? and (2) How should test items be constructed so that they will be homogeneous representatives of the test-item domain circumscribed by an objective?

SELECTION OF OBJECTIVES

Within each subject there are typically topics to be understood, intellectual skills to be acquired, constructs to be mastered, or dispositions to be promoted. In the revised IOX collections of objectives thus far completed, three levels of topics and skills have been labeled. The most general of these have been referred to as (1) *major categories*. Examples of major categories include "discrimination" (in the decoding collection), "sentence components" (in the transformational grammar collection), and "sets" (in a mathematics collection). Discrimination between "auditory sounds," "word order," and "cardinality" are examples of an intermediate level of skills referred to in the objectives collections as (2) *content general objectives*. The most precise statement of the skill is the specific (3) *objective* itself.

The individual who constructs IOX objectives-based tests has to make selections at all three levels of generality; namely, the major category to be included, the content general objective to be represented, and finally the specific objective to serve as the basis for the amplified objective (a concept to be discussed later). Note that we will use *titles* (a word or phrase) to serve as short form descriptors of each of these three categories. The actual title of one of the one-page spirit master tests should, of course, be as descriptive as possible while being quite brief.

Criteria for Determining Which Major Categories and Content General Objectives Are to Be Represented

In the total collection of tests for any one subject area it is likely that all significant categories will have some representation. However, practical limitations require that not all content general objectives will be represented. Typically a test developer will start by mapping out the eligible contenders for major categories in the subject field at hand. Much formal and informal advice from subject matter specialists should be secured at this point. Then having selected the major categories which will be measured by any particular box of 40–50 spirit masters (typically constituting one of the units we will distribute), the next job is to identify appropriate content general objectives.

The following criteria are offered as guidance (1) in determining how to sample a given major category and (2) in selecting content general objectives for those major categories which are selected.

1. *Importance.* What topics, what skills, etc., will be viewed by educators as the most important for that subject? We do not want to promote divergence here (often the IOX collections do just that); this is the time to opt for the consensus-derived winners. One way of getting at the importance of a set of skills is to try to identify those which constitute the skeleton of the subject, the really critical elements to which other elements can be added. Test constructors unfamiliar with the national picture in a given subject should inspect a wide range of curricular guided test books, etc. We do not want to be parochial in any sense.
2. *Economy of Production.* What topics, content, skills, etc., can we translate into tests rather readily? That is, if there are a dozen content general objectives in U.S. History which are generally conceded to be important by historians and history teachers, but we have an extant IOX collection which has rather good objectives and six test items per objective for seven of the twelve content general objectives, then we might lean toward those seven (other factors being equal).
3. *Practical Scorability.* Which major categories and content general objectives associated with them are *apt* to yield specific objectives and resulting test items which will be readily scorable? A readily scorable test item is not necessarily an objectively scorable item as in a multiple choice test. For instance, there are some topics in a subject which will require learner responses so complex in nature that we will be hard put to devise reliable scoring schemes. This does not mean that we should cleave to inconsequentials. Quite the contrary. We should be tapping the most significant kinds of learner behaviors we can. But given a pair of somewhat *comparable* major categories or content general objectives, we should opt for the more practicably measurable.

In some ways the selection of major categories for the tests will be relatively simple in contrast to the technical difficulties of selecting either content general objectives or specific objectives. Yet this represents an inordinately important decision. IOX will be distributing objective-referenced tests based on these major category selections, e.g., punctuation skills, sets, and quantum optics. Many educators will make their critical decisions regarding whether to use the IOX measures largely on the basis of the titles stemming from these major category selections. Therefore, they should be winning titles.

Similarly, having selected the major categories the selection of the content general objectives will be highly important, for such objectives will represent the next level of scrutiny for potential users of the tests. In *both* of these selection operations, the three criteria given above should be diligently employed.

An important question arises concerning the number of tests which should be selected for a given subject field. The selection of the spirit master method of duplication means that we will probably distribute boxes of 40–50 masters at one time. For instance, we are splitting the reading area into two boxes, decoding skills and comprehension skills, each with about 40 masters per box. Because we will have one box for form A and a second for form B, this means that a school can buy the four boxes and be set up to run off pretests and posttests to cover most of the critical reading skills. We will split up mathematics into several boxes of 40–50 one-page tests. New subject fields will have to be organized for test development purposes so that they will divide reasonably well into one or more of these 40–50 master boxes. It will be better, in general, to favor using fewer boxes since this will reduce the costs to school people wishing to employ our objectives-based measures.

Selecting Specific Objectives

In most domains more or less delimited by a content general objective there is more than one sort of specific learner behavior which might be used to reflect mastery of the domain. To illustrate at a rudimentary level, we might think of numerous ways that a learner could demonstrate mastery of basic subtraction operations involving two single numbers or their equivalent symbols. This means that the learner might be presented with test items such as X_1, X_2, X_3 or X_4 as shown below.

$\begin{array}{r}3\\-1\\\hline ?\end{array}$	$3 - 1 = ?$	$000 - 0 = ?$	Three less one is ?
X_1	X_2	X_3	X_4

We might encounter content general domains which could legitimately be tested by 20, 30, or even more such types of test items, i.e., specific objectives. Since we have decided to generate tests consisting of only *one* such type of test item, as opposed to sampling[1] from many or all of the eligible types of test items, which objective should be chosen? As will be depicted in the next section of this paper, the particular form of learner behavior selected will constitute the amplified objective which will guide the generation of our tests. Hence, the significance of the specific objective selection is considerable.

---

[1] A common practice in many standardized achievement tests.

The following criteria should be applied in deciding on the type of learner behavior which will serve as the specific objective, thereafter to guide the test construction:

1. *Widely Accepted.* The objective selected should be the most widely accepted as important by those in the field. Unlike the IOX objectives collections where we present a wide array of alternatives and then encourage educators to choose among them, here we will have to cater to the majority preference.
2. *Transferability Within Domain.* The form of learner behavior selected should be the most generalizable of those represented in the content general domain, i.e., a learner mastering the designated behavior requirements would likely be able to transfer that mastery to most, if not all, of the other eligible behavioral requirements in the content general domain.

   In making such a selection it is important to consider the entire range of learner behaviors with which we are concerned, i.e., both test-like events and real-world events. For instance, in surveying an individual's mathematical competence we should be attentive not only to the "Xs" which we can represent via standard test formats but to the "Xs" which might reflect such skills as the ability to make change in a supermarket or to complete one's annual income tax report.
3. *Terminality.* If there is a degree of possible hierarchy present in the contending types of learner behaviors under consideration, such that some are considered precursive or enroute to others, the chosen specific objective should represent the most terminal learner behavior. The learner behaviors should be those that will require a reasonable investment of instructional time. They should, in general, not be something which can be taught in a very few minutes.
4. *Transferability Outside the Domain.* Another consideration in selecting a specific objective is the degree to which that behavior, once mastered, will be transferable outside the content general domain, for example, to domains which might be learned by students in the future. For instance, certain skills acquired by students in one course (such as the ability to distinguish between fact and opinion) may have reference to many other courses. Such high transfer skills and intellectual constructs should be given high priority in the selection of specific objectives.
5. *Ease of Scorability.* In an effort to produce tests which have considerable practical utility, we must try to select learner behaviors which, other factors being equal, can be easily scored by those educators

employing them. Again, this does not limit us to selected response items, for in some instances we will surely find it necessary to utilize constructed response formats. (This may help distinguish the IOX tests from typical standardized tests.) Nevertheless, scoring practicality is a nontrivial consideration.

6. *Amenability to Instruction.* If there are certain intellectual potentialities which are relatively resistant to instruction, for example, native intelligence or cultural experience-based skills, then the teacher's efforts will probably not prove too successful. We should try to avoid these behaviors, selecting instead those that have a reasonable chance of being achieved through the teacher's instructional program.

Now how should these six criteria be employed in selecting the specific objectives? Since no handy scheme is available for mechanical translation into decisions, test constructors will have to be self-consciously attentive to each of the points. If the test developer has exhausted all rational alternatives, an arbitrary selection is always possible.

ITEM GENERATION

What is being proposed is far less ambitious than Wells Hively's system of domain-referenced testing—complete with item forms, item shells, replacement matrices, etc. Yet, the approach is derivative from Hively's strategy in that it attempts to set certain limits regarding the kinds of test items which will be constructed. Because Hively's system has, for some, proved too sophisticated for sustained use, the constraints to be employed are more modest.

Since a typical measurable objective often leaves too much latitude to those who must devise test items to assess it, we obviously need more delimitations than the run-of-the-Mager objective. What is suggested, therefore, is an *amplified objective,* i.e., an expanded objective which contains sufficient details regarding the nature of the measurement procedure to help an item-writer produce homogeneous items.

There are two major elements in an amplified objective: first, a delimitation of the *stimulus elements* and, second, a description of *learner response options.* Following each amplified objective, a sample test item must be provided to clarify further the nature of the amplified objective (although one hopes such additional clarification would not be requisite). The sample item *will not* be included in the test manual distributed along with the new IOX tests, since its main use is in connection with item generation. Amplified objectives, however, *will* be included in such manuals.

Stimulus Elements

The first thing the amplified objective (AO) must possess is a thoroughgoing description of what stimuli can constitute the test item. In one sense, of course, we might think of the test item in its entirety as constituting a stimulus for the learner, but for our purposes we will divide the item along classic lines as illustrated here:

Stimulus Elements	*Response Elements*
1. Mary had a little lizard.	True or False
2. How tall was Tiny Tim?	______ feet, ______ inches
3. The first U.S. President was:	a. Martha Washington b. George Gobel c. Martha Gobel d. None of the above

In considering the nature of the stimulus elements (left column above), it is necessary to spell out as much detail as seems needed (this will obviously vary from subject to subject, AO to AO) to reduce possible ambiguity regarding what to include in the stimulus section of an item. For instance, the following would be insufficient detail:

> When presented with a series of true or false statements regarding this nation's relations with Cuba, the learner will correctly identify those which are true.

The item writer does not have any guidance regarding what kinds of statements concerning Cuba-U.S. relationships are fair game. A better AO for such an objective might be put together something like this:

> When presented with a series of the following types of statements concerning U.S.-Cuba relationships, the learner will correctly identify those which are true:
>
> a. *Economic:* dealing with size of mutual imports of tobacco, rice sugar, wheat for the period 1925–1955.
> b. *Political:* dealing with status of formal diplomatic relationships from 1925 to the present.
> c. *Military:* dealing with the post-Castro period emphasizing the Bay of Pigs incident and the USSR missile crisis.

Now there is still lots of slack in the above AO, and we might tighten it up further by detailing the kinds of true or false items to be used. For instance, we could indicate that the forms of false items would involve *only* (a) switching the roles of the U.S. and Cuba, (b) distorting the chronology of events, or (c) using names of key individuals other than those actually included.

Even this simple true/false example illustrates how difficult it is to build sufficient constraints in the AO so as to limit meaningfully the set of eligible test items without, at the same time, trivializing that set of items. A test constructor may yearn for the simplicity of mathematics where an amplified objective such as the following does the job economically, yet without a major loss of significance:

> When presented with any pair of numbers, either triple-, double-, or single-digit numbers in vertical format, the learner will be able to supply the correct quantity indicated by the operation sign presented, i.e., (÷), (-), (×), (+).

Having generated the above AO, the item writer could then confidently dash off a series of test items such as these:

(a)	(b)	(c)	(d)
14	182	44	172
+ 2	× 13	÷ 22	- 14

Of course, even here there are problems. In items such as (c) above, do the division operations have to yield an even quotient? In items such as (d), is the top number always larger than the bottom? In such anticipatable instances, the AO writer will probably have to footnote the AO to indicate, for such cases, what is acceptable. Now some of these constraints will be relatively arbitrary while others will be quite important. The two chief factors guiding the AO generator, who will usually also be generating the subsequent test items, is (1) *economy of description* (will it take pages and pages of limitations or only a fairly detailed footnote?) and (2) *ambiguity reduction* (does the additional footnote delimitation markedly reduce the heterogeneity of items which might emerge from the AO in its present form?).

In working on an AO, IOX test constructors should take a middle position between the polar extremes of (a) sufficient detail for complete homogeneity of resulting test items and (b) economy of resource investment. The former would result in an extremely extensive AO statement, while the latter would yield a much more brief, easy-to-construct AO. It is clear that the kind of AO being recommended here will *not* delimit all of the possible test items, but it should markedly reduce the ambiguity associated with the objective. This is particularly important in that those educators who will be using our tests must be able to examine the amplified objectives we present in the test manuals, then judge whether the pool of items in the test (usually five or ten) are accurate indicators of that AO. Putting it another way, users must be able to judge whether the items are suitable representatives of the hypothetical pool of items circumscribed by the AO.

Learner Response Options

There are two ways a learner will be able to respond to the stimuli in a measurement situation; by *selecting* among choices presented by the test constructor or by *constructing* his/her own response. If the test items are to involve selected responses such as true/false, right/wrong, etc., there is little problem. If, on the other hand, the selected responses are from a multiple choice format, then the AO constructor must cite not only the characteristics of the elements which *should* be selected (the right answer), but also those elements which *should not* be selected (sometimes referred to as distractors). *Both* of these must be given in the AO. To illustrate, examine the following AO's delimitations regarding learner responses.

> When presented with one-sentence definitions of standard measurement operations or constructs* plus four possible correct answers, the wrong answers will always be drawn from the same set of constructs or operations.
>
> *Consisting of validity, reliability, objectivity, internal consistency, discrimination indices, distractors, negative discriminators, positive discriminators, norm-referenced, and criterion-referenced.

In the case of constructed responses, it is imperative that the AO generator must supply criteria by which to judge the adequacy of the learner's constructed answer. For instance, if an essay answer is to be given, then what features *must* the essay incorporate in order to be considered an acceptable answer? These features, preferably with good and bad examples, must be supplied or the judgment of responses becomes guesswork.

Item Format.

A sample item (or a sample structure for all items) must be produced by each AO generator. These will *not* be supplied in the test manual, but will prove useful to those IOX staff members who must review AO statements *prior* to the production of sufficient test items, to the test writers, as well as to the AO generator him/herself.

Thus, we can foresee each AO consisting of an augmented set of test item delimitations, perhaps a half-page or so in length, sometimes with further delimiting footnotes. After this, a sample item or item format will be supplied.

Efforts to tie down completely all of the requisite elements needed to produce an exemplary amplified objective have, thus far, proved unrewarding. The particular requirements arising from the specific content and behavior considerations occurring in differing objectives seem to work against tidy checklists.

DEVELOPMENT SEQUENCE

There are typically advantages associated with being able to see all the aspects of a development operation in order to note how each of the parts relates to the whole. Accordingly, a short description will be presented of the events involved in the development of an IOX objectives-based test. Following the schematic diagram (Figure 1), which will serve as an overview, each of the steps in the operation will be briefly detailed.

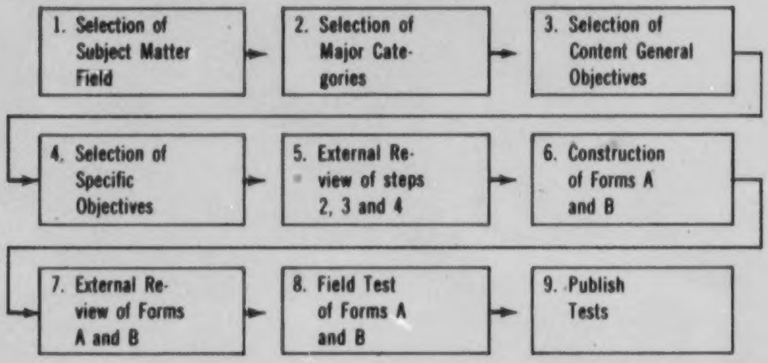

1. Selection of Subject Matter Field
2. Selection of Major Categories
3. Selection of Content General Objectives
4. Selection of Specific Objectives
5. External Review of steps 2, 3 and 4
6. Construction of Forms A and B
7. External Review of Forms A and B
8. Field Test of Forms A and B
9. Publish Tests

Figure 1. Steps in the Development of IOX Tests

Before proceeding to an examination of the nine steps depicted in Figure 1 it is important to note that while no formal identification of "review-revise" operations has been given, at this point, these would occur between almost every step in the enterprise. Sometimes the review will be very informal, supplied by a co-worker assigned to that test development project, for example, between Steps 2 and 3. Sometimes there will be very specific internal review operations. These will be described in the next section of this paper. But for overview purposes, let us turn to an examination of the nine steps identified in Figure 1.

1. *Selection of Subject Matter Field.* The first step in the development of a test is the decision as to what subject matter area we will be working in. At the outset, it is apparent that we are working in the most commonly taught fields, that is, reading, math, social science, and science. In some instances these selections were made based on the capabilities of existing staff. In other instances, they were simply the obvious high-need choices. Generally speaking, the selection of a subject matter field is a decision made at the directors' level based on existing resources, e.g., office space, financial support, and personnel.
2. *Selection of Major Categories.* Once the staff member or staff members have been assigned to the test construction activity resulting from Step 1, the next task is the selection of the major categories which will be treated in the test. These rather large "chunks" of content (e.g., skills, attitudes, etc.) will subsequently be represented either by themselves, or perhaps in concert with other major categories, as IOX tests. The criteria for the selection of these major categories were supplied earlier.
3. *Selection of Content General Objectives.* It is at this point that the test constructors are beginning to develop a map of the elements

to be measured in the tests. Criteria for the selection of content general objectives were described earlier. Here the test constructor should be getting some reaction from other people knowledgable in the field (friends or associates). This does not constitute a formal external review, but informal reactions from others at this point should be most helpful.

4. *Selection of Specific Objectives.* Having decided upon the content general objectives, the next task of the test constructor is to develop amplified objectives which are essentially expanded statements of the instructional objectives (such that they will more appropriately delimit resulting test items). Conceding that these amplified objective statements will be modified as the actual test items are subsequently prepared, the test developer should produce at least one sample item and move as close to closure as possible regarding the structure of all amplified objectives. It is at this point that the developer really has to cope with the 40–50 masters box limitation and split up, or collapse, the subject field accordingly.
5. *External Review of Steps 2, 3, and 4.* This is the juncture at which we need to bring in some person from the field, typically a teacher who is involved in teaching the particular subject involved, to survey Steps 2 through 4. This probably will be a half-day or full-day assignment when the teacher meets with the staff working on the particular test. We obviously have to select external reviewers very carefully for this assignment. On the basis of their reaction, there will probably be some changes made in the map of major categories and content general objectives.
6. *Construction of Forms A and B.* It is now time to actually produce Forms A and B of the test. Any departures from format guidelines should be approved in advance. A copy of the correct answers, which accompany the test as it is reviewed, should be retained by the developer.
7. *External Review of Forms A and B.* Once more we need to get some practicing educator to examine the actual tests we will be distributing prior to their field trial. We can probably locate some individuals who will do this on a test-by-test remuneration basis. We simply have to get someone who is out there in reality looking at our tests to detect gross procedural or content flaws. Often this can be the same individual (typically a teacher) who was involved in Step 5.
8. *Field Test of Forms A and B.* At this point we have to try out the tests with a limited number of students (for example, 5–10) to get a fix on elements in the instrument which may be drastically wrong. It is recommended that members of our staff *other* than those individuals who have developed the tests go out to administer the test, but that a member of the construction team accompany this administrator so that if procedural or content questions arise during the

testing they can be treated by the individual most conversant with the instrument's development.

9. *Publish Tests*. Assuming that the tests will be revised on the basis of the field tests, we still need to have someone check them over one last time before we prepare the copy to take to the printer. After this is done, the next step is to follow routine production procedures prior to first publication.

INTERNAL REVIEW

The rationale for test development in our objectives-based tests is quite different from that of conventional test publishers. We will be relying far less, particularly at the outset, on results from extensive field tests with their resulting reliability and validity coefficients. Instead, our general strategy will be to emphasize the content validity of our tests via a priori judgments regarding the congruence between the test items and the amplified objectives which they are designed to measure. Accordingly, we need to set up specific check points where we can be sure that such congruence exists. Referring to the previous description of steps in the development sequence, we see the addition of points Ⓐ, Ⓑ, and Ⓒ which are intended to represent three separate internal review operations in Figure 2 below.

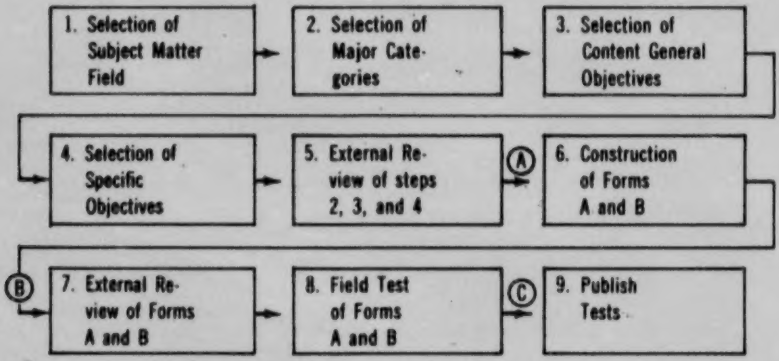

1. Selection of Subject Matter Field
2. Selection of Major Categories
3. Selection of Content General Objectives
4. Selection of Specific Objectives
5. External Review of steps 2, 3, and 4
Ⓐ
6. Construction of Forms A and B
Ⓑ
7. External Review of Forms A and B
8. Field Test of Forms A and B
Ⓒ
9. Publish Tests

Figure 2. Internal Review Points of IOX Tests

At point *A* our internal reviewer will examine all amplified objectives (AO) and the one (or more) sample item(s) accompanying each AO. Sometimes the test developer has produced all form A items at this point. The focus of this review will be to sharpen the AO if necessary.

At point *B* the intent of the review will be to see whether *all* items in the two forms of the test match the AO as well as to make a cursory check for any glaring deficiencies in the tests themselves, e.g., format. Thus there are two signoffs at this point, one for congruence and the other for format. Review *B* will probably be the most significant of the internal reviews.

At point *C* the review will be relatively brief, providing a last look at the tests which will be dependent on how many changes, if any, have been made as a result of review B, or the field test (step 8).

A signoff sheet will be used to accompany all tests as they move through steps 1–8. Each test should be placed in a manila folder to which the signoff sheet is attached. This will permit us to keep track of the status of a given test. A master status sheet will be used to monitor the progress of a test as it passes through the various development stages.

UNRESOLVED PROBLEMS

The foregoing remarks may at times have conveyed a tone of authority not consonant with the writer's confidence in the state of the technology. A good many of the recommendations for our project involving the production of IOX objectives-based tests are little more than experientially derived guesswork. Prominent among the problems we face are the following, arranged in order of decreasing significance to us:

1. What techniques can be devised which will permit objectives-based test developers to improve their instruments on the basis of empirical tryouts in the same ways that conventional test developers have been doing for years (e.g., total test reliability, item reliability, item homogeneity, objective-item congruence)?
2. How can a replicable set of guidelines be produced which will allow one to economically yet definitively constrain item-writers who will produce objectives-based tests?
3. Are there technical rules which can be produced to aid reviewers in judging the congruence between test items and the objectives on which they are based?
4. Can a technology be devised to assist objectives-based test designers to delineate satisfactory criteria so that items calling for constructed learner responses can be employed with the expectation that the resulting responses can be reliably scored?
5. Was our decision defensible to devise tests which assess only certain specific objectives (for example, $\times_4$) versus sampling from many objectives ($\times_1$, $\times_2$, $\times_3$, $\times_4$, etc.)?

If problems such as the above can be addressed with a view to supplying some *practical* resolutions, the increasing numbers of individuals who will be constructing objectives-based tests will obviously be advantaged. More importantly, the ultimate beneficiaries will be the learners who will then be assessed with such tests.

A JUDGMENTAL APPROACH TO CRITERION-REFERENCED TESTING

H.A. Wilson
National Assessment of Educational Progress

Practical applications and experimentation in science are grounded in a solid theoretical basis; intellectual activities lacking such a solid theoretical basis might realistically be considered arts rather than sciences. Much of education is more art than science in that many educational activities are based on the intuition and judgment of practitioners rather than on logical extensions of quantifiable theory. The field of educational measurement is an exception, however, since an extensive body of theory has been developed that guides the activities of norm-referenced testing. However, in educational measurement there is a much older tradition that attempts to determine the absolute achievement of the individual or population without regard to interpersonal comparisons. That tradition, which is currently called criterion-referenced testing, can call on much of the statistical technic that is used in other fields of measurement. Criterion-referenced testing, however, faces some important problems in basic theory that norm-referenced testing can, by definition, safely ignore.

Activity in criterion-referenced testing, as with other considerations in education, cannot be delayed until basic applicable theory is developed. Schools cannot close their doors until a comprehensive theory of learning is found. Neither can assessment activities be halted. We must, instead, rely on human judgment to solve practical problems while we work on basic theory. This is the situation currently faced by the National Assessment of Educational Progress (NAEP).

A brief overview of the history and purposes of the National Assessment might be useful as background for a discussion of NAEP's responses to important theoretical questions in the area of criterion-referenced testing.[1]

By the early 1960's many billions of dollars were being invested annually in the formal education of our young people. The only available measures of educational quality resulting from this investment had been based upon *inputs* into the educational system such as teacher-student ratios, number

---

Preparation of this paper was supported by funds from the National Center for Educational Statistics, U.S. Office of Education, Department of Health, Education, and Welfare. Points of view expressed or opinions stated do not necessarily reflect official USOE position or policy.

[1] This description was adapted from F. Womer, *What is National Assessment* and from C.J. Finley and F.S. Berdie, *The National Assessment Approach to Exercise Development* (Denver: National Assessment of Educational Progress, Public Information Department.)

of classrooms, and number of dollars spent per student. The tenuous assumption had been that the quality of educational *outcomes*—what students actually learn—was directly related to the quality of the inputs into the educational system. No significant direct assessment of educational outcomes had been made. The typical state-administered or school-administered achievement tests, which provided scores whereby one student could be compared with others, were useful for categorizing students; but they provided very little information about what students were actually learning.

This insufficiency of information became the concern of Francis Keppel, United States Commissioner of Education (1962–1965), who initiated a series of conferences to find ways in which it might be overcome. In 1964, as a result of these conferences, John W. Gardner, president of the Carnegie Corporation, asked a distinguished group of educators and lay persons to form the Exploratory Committee on Assessing the Progress of Education (ECAPE). This committee, chaired by Ralph W. Tyler, was to examine the possibility of conducting an assessment of educational attainments on a national basis.

After much study, ECAPE deemed that it was feasible to assess the knowledges, understandings, skills, and attitudes in 10 subject areas (Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies, and Writing) at four age levels (9, 13, 17, and adult—ages 26–35). The project began its first assessment of the subject areas of Science, Citizenship, and Writing in Spring of 1969. Later that same year, the project came under the auspices of the Education Commission of the States and was named the National Assessment of Educational Progress (NAEP).

For the first time, there would be a *direct* measure of educational outcomes which could be utilized by school systems to improve the educational process. Since NAEP is to be an ongoing project, it will eventually be able to assess changes in these knowledges, understandings, skills, and attitudes to determine any changes in educational outcomes.

Many people prominent in education and measurement have contributed heavily to the purposes and processes of NAEP. A brief and very incomplete roster would include, besides Tyler, Francis Keppel, John W. Gardner, Jack Merwin, Frank Womer, Stanley Ahmann, John Tukey, Frederick Mosteller, and Lee Cronbach.

Two subject areas are currently being assessed each year with a five-year cycle for reassessment within a given subject area. The five-year assessment-reassessment cycle and the 210 minutes allotted to each subject area at each age level in an assessment year place very practical constraints on the design and production of exercises (test items). The five-year cycle requires continuous exercise development effort and limits experimental and validation activities. The time allotment limits the number of exercises administered and, hence, the depth of coverage for each objective.

UNIVERSE DEFINITION

Some of the most intriguing questions in the field of criterion-referenced measurement have to do with the rigorous definition of a domain of reference (subject matter) and of a universe of behaviors within that domain. This paper will briefly summarize some of these questions and indicate the general thrust of NAEP's responses. The responses discussed should be viewed as current positions of NAEP regarding the basic problems; they are in no sense offered as definitive solutions.

Two questions must be asked:

1. What constitutes a definition of a domain of reference or a universe of behaviors?
2. When can we be sure that a complete definition is achieved?

Since the problems of defining a domain of reference and a universe of behavior are parallel, discussion of a domain of reference can serve as a model for discussion of a universe of behaviors.

It is clear that a complete definition of a domain of reference must include all knowledge, skills, and attitudes directly related to the subject area and exclude all those that are not related. A similar statement could be made for defining a universe of behaviors by substituting "behaviors" for "knowledge, skills, and attitudes." Such a definition need not be an enumeration. Indeed, such an enumeration would be useless because of its extensive, if not infinite, length.

What is needed, then, is a method of statement generation that will produce relevant, and only relevant, statements. We can be sure that a complete definition is achieved only when it can be logically shown that any statement or question stemming from our statement-generation mechanism is or is not a member of the set of questions and statements contained in that domain or universe.

Without a logically complete knowledge generator it is not possible to make statistically defensible and generalizable statements relating individual or group performance to a subject area by means of a restricted set of items. Without a complete definition of the domain of reference and a universe of behaviors, all statements about the results of a criterion-referenced test must be confined to the items in that test without further generalizations. Clearly, this is not the purpose of any test maker.

Several approaches to the problem of generalizability can be found in the literature. One approach is to ignore the problem altogether. Another is to indicate how certain domains and universes can be defined and systematically sampled. Typically, however, those domains and universes that have been discussed are narrowly restrictive or trivial or both. For example, tests of knowledge of word meanings can be constructed by defining the domain of reference as the Merriam-Webster Collegiate

Dictionary, 7th Edition. All statements about words included in that dictionary are relevant and all statements about words excluded from that dictionary are not relevant. One can then define the universe of behaviors as responses to a cloze test on the definitional entry for each word. Many schemes can then be devised for systematically sampling both the domain of reference and the universe of behaviors. Item generation rules can be devised which will produce any number of equivalent tests and the results of these tests can indeed be generalized to knowledge of word meanings as defined in the domain of reference. Such schemes are of little value, however, in constructing tests to assess knowledge, skills, and attitudes in broad areas such as social studies, literature, music, or art.

OBJECTIVES

It is clearly beyond the current state of the art to define the universe of behaviors for a complex area in the strict sense discussed above. Yet it is equally clear that a set of exercises (test items) which form a coherent assessment of a subject area cannot be constructed without some definition of the domain to be tested. In response to this conundrum NAEP has taken a judgmental rather than a statistical approach to universe definition.

The term "judgmental" is used to indicate reliance on human judgment rather than logical or statistical proof. We define our universe by producing a set of objectives that represents a consensus of opinion covering many segments of our society regarding the important goals and outcomes of our educational processes in respect to a given subject area. But the question might well be raised, "Why add yet another formulation of educational goals and objectives to the already existing plethora of such documents?" It is certainly a reasonable question and yet one that is easily answered in terms of NAEP's mission. NAEP, as its name states, is a *national* assessment and as such is compelled to attend to those aspects of education whose definition and evaluation can be agreed upon for the society as a whole. Most of the myriad statements of objectives are currently produced by and for the use of schools at the local and state level. NAEP must go beyond that restricted viewpoint to identify goals that are accepted nationally.

Since NAEP is also an assessment of change in educational outcomes over time we have the further responsibility to examine and revise our codifications of objectives on a systematic, cyclical basis. The twin requirements of demonstrable national significance and continuous revision justify the effort to produce statements of goals and objectives that are unique to our own needs and purposes.

NAEP defines the domain of reference in a subject area by arriving at a national consensus statement of goals in that area. Goals are stated in the form of overall objectives with attendant levels of sub-objectives. The form and structure of the objectives vary from one subject area to another and between assessment cycles within a single subject area. For example, a major objective and its sub-objectives for cycle 1 of Music were stated as follows:

III. LISTEN TO MUSIC WITH UNDERSTANDING.

A. *Perceive the various elements of music, such as timbre, rhythm, melody and harmony, and texture.*

1. Identify timbres.

Age 9	Identify by categories the manner in which the instruments are played (e.g., struck, bowed). Identify individual instrumental timbres—unaccompanied. Identify individual instrumental timbres—with accompaniment.
Age 13	(in addition to Age 9) Identify individual vocal timbres—with accompaniment. Identify ensemble timbres, instrumental and vocal.
Age 17 Adults	Identify by categories families of related timbres (e.g. woodwinds, plucked strings). Identify individual instrumental timbres—unaccompanied. Identify individual instrumental and vocal timbres—with accompaniment. Identify ensemble timbres, instrumental and vocal.

A much more loosely defined objectives structure was produced for the first cycle of Literature assessment as shown by the following example:

III. DEVELOP A CONTINUING INTEREST AND PARTICIPATION IN LITERATURE AND THE LITERARY EXPERIENCE

This goal is directed at assessing interests and attitudes; for the most part the goal is relevant to Age 17 and Adult.

A. *Be intellectually oriented to literature.*

This goal asks of the individual a recognition of the importance of literature to the individual and society, and a recognition that literary expression requires a number of forms to enable it to become an art.

All ages	Recognize the importance of literature to an understanding of cultures distant in time or distinct in history. Recognize the importance of literature to a comprehension of the diversity and homogeneity of man. Recognize that participating in the literary experience is a prime form of enjoyment.
Age 17 Adults	Recognize the necessity of a free literature in a free society. Recognize that the art of literature involves a close connection between form and content.

The process of identifying and explicating objectives or revising those used in the previous cycle of assessment of a subject area is somewhat complex and occupies a time span of approximately nine months. A search of recent literature is made to identify new trends in the subject area. The literature search is coupled with an examination of existing sets of written objectives such as those brought together by the Instructional Objectives Exchange. This material forms a background for a number of working and review panels that produce and refine the objectives to be used as the basis for exercise development and for reporting of assessment results.

In the early years of NAEP, objectives development was done by sub-contractors (American Institutes for Research, Educational Testing Service, Science Research Associates, etc.). They studied the literature, examined existing objectives, and produced a document that was critiqued by a variety of consultants and then revised. This plan was followed for the objectives development of most of the first cycle assessments. Leaving objectives development in the hands of the contractors who then wrote the exercises not only produced objectives of uneven quality but was also liable to produce only those objectives that were most easily measured while neglecting those that are difficult to measure but still important to the education community.

With these considerations in mind, the task of producing objectives was no longer left to sub-contractors but was made part of the direct responsibility of the Exercise Development department of NAEP. The standardized procedure that is now followed for developing objectives begins with a mail review by subject matter experts of the objectives from the previous cycle. This mail review is followed by a conference in which consultants determine the broad outlines of the desired revision. A sub-set of consultants from the first review conference produces a draft of the revised objectives following the guidelines stemming from the conference. This draft is reviewed by mail by members of the first conference and a second draft is produced based on the resulting comments. This draft is then reviewed by a second conference of consultants, some of whom were present at the first revision conference. Consensus among the consultants is reached on the remaining points at issue and the document is adjusted accordingly and given final editorial polish.

The working and review panels are composed of consultants drawn from two major groups: scholars and educators within the subject area and qualified and interested laymen. Between 35 and 50 consultants are involved at one time or another in the development of objectives. Consultants are chosen with serious attention to representation by region (northeast, southeast, central, and west); type of institution (university, four-year college, junior college, secondary and elementary schools, and private schools); and race and sex. Wherever there are clearly defined schools of thought that hold differing positions in a subject area, care is taken to assure representation of each of the conflicting points of view. The above procedure describes the selection of consultants who serve actively on panels; in the case of mail reviews, a much larger number of people are involved.

Though far from ideal, the method just described does produce a set of objectives that represents as nearly as possible, within the constraints of time and money, a national consensus on educational goals and objectives currently valued by our society. Great emphasis is placed on producing objectives that are important without regard to their measurability. NAEP views the objectives as defining the broad domain within which exercises are to be written and as a mandate from our society to produce data on related educational outcomes.

There are a number of important questions still unresolved in the area of objective development. The question with the deepest theoretical implications is, "To what level of sub-objective and age-specific behavior should objectives be taken?" The major objectives are generally few in number and are of such general nature that they provide only an ambiguous guide to exercise development. At each level of sub-objective the domain of reference is more clearly defined but how clear that definition can be or should be is still an open question. There is currently a large variation in this matter from subject area to subject area and between assessment cycles within any given subject area. The use of age-specific behaviors in the objectives furnishes the clearest definition and guide for exercise development. However, it is again a question of whether to view age-specific behaviors as an exhaustive list of all possible behaviors (an obviously impossible task) or to view them simply as guidelines and illustrations for the exercise developers.

A second and related question has to do with the feasibility of developing some sort of hierarchical scheme of cognitive and affective objectives. Many such schemes have been devised but the question remains whether it is possible or even advisable to choose one plan to the exclusion of all others.

A final question has to do with standardizing the format of objectives. It has been suggested that from a quality control standpoint, a standardized format and framework of objectives should be developed and applied to all subject areas. There is no solid agreement, however, that this plan, if it could be implemented, would be desirable. The discussion on this point revolves around the issue of the amount of freedom to be allowed to the developers of the objectives to express in their own way those aspects of the subject area that they feel to be most important in our educational scheme.

ITEM GENERATION RULES

In criterion-referenced testing it would be desirable to identify a generally acceptable method for item construction. In the strict sense, such a method should provide a systematic sampling of a previously defined universe of behaviors. Further, the method should provide a set of rules which, if followed by more than one person or group of item writers with equivalent knowledge, would produce equivalent tests. We have already discussed the difficulties involved in domain and universe definition in the complex areas that are of interest to NAEP. Since the universe of behaviors has not been well defined, a systematic sampling scheme is difficult to devise. When we examine the notion that a set of rules may be clearly enough stated that equivalent tests may be generated from them, it is easily seen that while such rules are useful in narrowly specialized areas, they are not definable in other more complex areas. Tests of arithmetic computational skills, tests of word meanings, and spelling tests have been constructed using such rule sets. Indeed, on occasion, rules have been embodied in computer programs which will generate equivalent tests ad infinitum. While complete in themselves, such tests unfortunately fall far short of being comprehensive tests of mathematics, reading, or writing.

Assuming for a moment that solutions were at hand for problems of defining the universe of behaviors and of stipulating an adequate set of rules for generating items, we are still faced with a question that has serious theoretical consequences. The question might be phrased as, "How much is enough?" "How many items are necessary to constitute an acceptable test of an objective?" If the objectives are complete through the identification of one or more levels of sub-objectives under each major objective, and if each sub-objective is adequately tested, then we can certainly claim that we have an adequate test of a major objective. However, such a plan simply puts off the problem to another level of detail. We are still faced with the central question of how many items are necessary to test the lowest level sub-objective or any given age-specific behavior.

NAEP EXERCISE DEVELOPMENT

In light of the problems outlined above, we may move to a brief discussion of the methods used by NAEP to generate exercises (test items). None of the activities described below are presented as final solutions but many of our item generating activities, while perhaps tangential to the central problems as stated above, do stem from our abiding concern for such problems. Again, as in the definition of domains of reference and universes of behavior, we continue to use a judgmental approach in the sense of relying primarily on the judgment of experts in the subject matter area.

Following the development of objectives, contracts are awarded through competitive bidding for the generation of exercises to assess the objectives. The amount of exercise material to be developed for each sub-objective is based on a "weighting" scheme. Weights are assigned by subject matter experts who are experienced with students at the four age levels. For example, the major objectives are weighted for their relative importance for nine-year-olds by teachers who have experience with that age group. Each sub-objective is then weighted for its relative importance within the major objective. This scheme is continued to the lowest level of sub-objective. The weights for an objective may differ widely over age groups reflecting the importance of that objective at one age as opposed to another.

The use of weights is in some sense a response to the problem of providing adequate coverage for each sub-objective. Since the weight of the sub-objective is an index of its importance in relation to other objectives at a given age level, such weights can easily be translated into percentages of the total assessment time that it would be reasonable to spend in assessing that particular sub-objective. Of course, this method of specifying coverage accounts only for amount of material related to its importance and does not speak to the issue of relating coverage to the complexity of the various objectives and sub-objectives.

NAEP has paid a great deal of attention to the problem of giving contractors an adequate framework for preparing the kinds of exercises that will achieve coverage through a variety of approaches. We have arrived at a general notion of exercise prototypes which are neither rules for exercise generation nor examples of specific exercises, but which rather attend to those aspects and variables of exercise generation that can be discussed. NAEP exercise prototypes are viewed as a tree structure showing mutually exclusive categories for four variables: administration mode, stimulus mode, response mode, and response category. The administration mode is dichotomous: an exercise can be administered either individually or to a group. Branching from the administration mode we define the stimulus mode as audio, visual, other senses (tactual, olfactory, etc.), or some combination of the three. From each stimulus mode we show a dichotomy of response alternatives or response mode: objective (multiple choice) and free response. Branching from each response mode, finally, we define response categories as written, verbal, role playing, group interaction, and other physical action.

Such a tree structure results in 80 (2 × 4 × 2 × 5) possible prototypes. It is clear that not all possible prototypes are applicable to any given subject area. A panel of subject matter experts selects those prototypes that are most reasonable for assessing a subject area. Their input, in conjunction with practical considerations of cost of administration and scoring, provides the specification of percentage ranges (minimum and maximum) in terms of minutes of material as guidelines for the contractor. The subject matter experts also produce exemplary exercises within the subject area for each prototype specified. The use of prototypes as a control for coverage through a variety of approaches is frankly experimental. Its first use will be in the current redevelopment of literature assessment, but it is expected to provide a more balanced body of exercises.

Working within the weighted objectives, prototypes, and exemplary exercises, contractors produce the specified minutes of exercise material for assessment of a subject area. Each exercise produced by the contractor must be accompanied by a rationale relating that exercise to the sub-objective that it purports to measure. It must also be accompanied by a rationale relating that exercise to other exercises within the body of material to be used in the assessment.

The exercises received from the contractor are subjected to reviews by each of three groups: the NAEP staff, subject matter experts (scholars and educators), and qualified laymen. In addition to reviewing the exercise itself, the rationale relating that exercise to a sub-objective and to other exercises in the body of material is also brought under scrutiny. Some exercises survive each review session; others are sent back to the contractor for suggested revisions and others, hopefully a small percentage, are rejected as being without merit and are no longer considered for use in the assessment.

Those exercises that have survived the reviews, either in their original or revised state, are then given a full field trial. Each exercise is tried out during its developmental stages by the contractor and is submitted to NAEP accompanied by data from three sub-units of the population: extreme inner city, extreme rural, and affluent suburb. Data from the developmental tryouts consist of timing information, overall percentage correct responses, percentages of responses for each foil in a multiple choice exercise, and the beginnings of a scoring guide or response categorization in the case of free response exercises. While these data are gathered from three sub-units of the population, the number of subjects contributing from each population is necessarily small. To increase the reliability of this sort of data, we run extensive field trials on a national sample. The field trials, while far less extensive than the actual assessment, are large enough to yield reliable data and at the same time they point up regional biases and administrative problems that might otherwise have been missed.

Following the field trials, the pool of exercises is reviewed by the Office of Education for possible offensiveness in sensitive areas. Exercises surviving this last review by USOE are then examined in selection conferences by successive panels of subject matter experts.

Since a precise attrition rate through all the reviews is unpredictable, we order a considerable overage of material from the contractor. This overage is on the order of 100% plus an additional 20% that allows for contractor creativity outside of the specifications and guidelines furnished by NAEP.

Since we are constrained to a total of 210 minutes of assessment for each subject area, a selection conference is necessary to choose the best from among surviving exercises. Consultants at the selection conference are required to pay close attention to maintaining the balance over objectives and sub-objectives that was specified in the original contract and to the relationships between exercises that form a coherent assessment.

VALIDITY

In terms of the assessment exercises, the two major concerns of NAEP are for their content validity and importance. Two questions are continually asked at every exercise review conference: "Is this exercise a valid measure of the objective for which it was written?" and "If it is valid, is it an important or a trivial measure of the objective?" Importance can only be established by following the judgment of subject matter experts. Human judgment is also the primary check on validity.

For some of the exercises, however, another measure of their validity can be obtained by examining the assessment response data. If an item is administered to two groups, one of which has had no training or experience in the area while the other has had extensive training, the results can be viewed as one measure of the item's validity. In the ideal case, a valid item would yield a score near zero for the untrained group and approach 100% correct for the highly trained group. Such a test is approximated for those NAEP exercises that overlap age groups. It may be assumed that seventeen-year-olds have had more training in a given subject area than thirteen-year-olds when training in that area is a continuous process. The same assumption may be made for comparison of thirteen-year-olds and nine-year-olds. If the same exercise is administered to the three age levels, an increasing percentage of correct responses from nine- to seventeen-year-olds can be accepted as some assurance of the item's validity. In general, such has been the case with NAEP data. If a contrary instance is found in the field trials, that item is examined closely. If an adequate explanation is not evident, the item is dropped from the assessment.

SUMMARY

Test construction is not the strictly logical process that we might wish it to be. This is particularly true in a large on-going project such as NAEP. Most of the really deep questions can only be answered by the exercise of well informed human judgment. Criterion-referenced testing is still a term in search of a definition. It has been suggested that NAEP's exercises might be more properly called "objectives-referenced" tests. That is a reasonable title for our efforts since we are attempting to assess the degree of achievement of stated goals without reference to a predetermined level or criterion. Whatever the appropriate title may be, we share the concerns of all workers in the field for the same basic questions. But until satisfactory scientific solutions have been found, we must rely on the best human judgment available.

MEASUREMENT CONSIDERATIONS IN INSTRUCTIONAL PRODUCT DEVELOPMENT

Robert L. Baker
Southwest Regional Laboratory for Educational Research and Development (SWRL)

The psychometric revolution that has been smoldering over the past decade and finally ignited in the "criterion-referenced test movement" will predictably spread throughout education during the next decade, and will generate consequences that go well beyond the boundaries of psychometry (Schutz, 1972). Even now it is obvious that concern with psychometric dogma reflected in such questions as "Is the criterion-referenced test just a special instance of the norm-referenced test?" and "How can the reliability of criterion-referenced tests be assessed?" is misplaced. Focusing on such questions is about as productive as the programmed instruction research of the 1960's related to overt-covert and large-step—small-step issues.

Recent instructional research and development has demonstrated that formal measurement can indeed fulfill important roles in producing instructional programs to meet prespecified objectives. However, full exploitation of this role requires control over non-psychometric as well as psychometric variables. The purely technical aspects of psychometry provide great capability for instructional product development. Conventional psychometric procedures can readily be adapted to generate measures which provide adequate bases for those instructional decisions that can currently be made and effected. Such practice, however, is not sufficient to advance the state-of-the-art for improving instructional effectiveness. The interface between psychometry and instructional development must include greater attention to instructional decision algorithms that are defined as functions of achievement measures anchored systematically to the manipulable conditions that produced the achievement. This consideration will encompass not only the specifications and development of instruction but also the installation and continuing operation of instruction. The effectiveness of specified instructional decision algorithms is dependent upon well defined assessment procedures that are easily reflected in defined behavioral classes of interest and anchored in manipulable instructional determinants.

The manipulable determinants of achievement in developing instructional programs are materials and procedures. To be useful in a development context tests must be designed and constructed in a manner that defines the explicit rules linking patterns of test performance to behavioral

---

Preparation of this paper was supported by funds from the U.S. Office of Education, Department of Health, Education, and Welfare. Points of view expressed or opinions stated do not necessarily represent official USOE position or policy.

referents anchored in sequenced instructional materials and procedures. Further, to be useful in an operating instructional context tests must be configured in such a way that a particular decision algorithm may be applied with little inconvenience.

The testing requirements following from these conditions are manifold, and the scientific and technological bases for getting the job done range from adequate to non-existent. However, absence of these bases cannot be permitted to halt development efforts. We must identify the immediately available resources for developing effective instruction and move as quickly as possible to completion of first-generation shelf items, recognizing that the items thus produced represent only a beginning of "more to come" from programmatic educational research and development currently in progress.

The remainder of the paper will view selected psychometric requirements and strategies as they interface with selected non-psychometric requirements of developing and delivering effective instruction. The view will be from within defined SWRL research and development activities and state-of-the-art capability.

MATTERS THAT ARE WELL WITHIN SWRL STATE-OF-THE-ART

The instructional development technology described in this section is readily available in shelf-item or easily adaptable form.

Writing Instructional Objectives

The "how-to" information for stating well-formed instructional objectives has been available for some time. A convenient recent synthesis of this information is contained in *Instructional Product Development* (Baker & Schutz, 1971). By reading this information, an interested high-school-graduate-equivalent person can acquire all of the information required to state well-formed objectives. However, the time-consuming and thought-challenging task of *what* outcomes to prepare remains to be done. But this is a matter of doing the job, rather than of not knowing *how*.

When the job of preparing well-formed instructional outcomes has been completed, one is at best at the *beginning* rather than at the end of instructional effectiveness. But the beginning is firm, rather than wishful.

Criterion-Referenced Test Construction

Not only does instructional development necessitate prespecified instructional outcomes, it also requires a means of assessing the attainment of these outcomes. This involves test construction activity. To be minimally useful the tests must be specifically referenced to a prespecified structure of achievement. To be maximally useful the tests must be specifically referenced to defined instructional materials. A consequent requirement is to define criterion behavior in the specification of the limits of a population of responses called for in the instruction which defines the criterion behavior rather than in a list of responses which exemplify it. This is not a new concept; it was encompassed by earlier discussions of content validity (Lennon, 1956) and Bruner's (1960) discussion of the structure of the subject matter knowledge. However, specific procedural cues for meeting the requirement were not available until Hively (1963) introduced the "item form."

The item form and related processes provide a neat system for blueprinting tests that meet all of the requirements of the psychometric concept of content validity and at the same time contribute to the definition of the behavioral structure of the subject matter domain treated. A collection of item forms sequentially ordered, together with the replacement sets for the variable elements, could adequately define a universe of content across specified outcome areas. When such procedures are more generally exploited the impracticality of constructing criterion-referenced tests for complex behavioral-content domains cited by Ebel (1971) is overcome.

Instructional Specifications

An item form defines classes of behavior, but it does not indicate how the behavior is to be established. However, as strings of item forms are prepared, it is possible to arrange them into tentative sequences that constitute an operational "cognitive map" of a subject matter useful in guiding both instructional and evaluational efforts.

The "instructional specifications" approach (Sullivan, Baker, & Schutz, 1971) provides a set of procedures for mapping out the instructional and assessment sequences consistent with the item form. The instructional specification (IS) is a convenient guide to the development of effective instruction for a given instructional objective. A well-constructed IS per instructional objective provides answers to the following questions:

1. What outcomes (objectives) will the successful learner attain as a result of the instruction?
2. What information (cue) will be given the learner to increase his ability to perform the desired behavior?
3. What procedures (mastery items) will be used to provide for practice and assessment of the desired behavior?
4. What are the characteristics (limits) of the correct responses or response choices for the desired behavior and what are the characteristics of plausible but incorrect responses?
5. What relevant skills (entry skills) must the learner possess prior to the instruction for the present objective?

Instructional programs that are developed properly from a set of written IS's incorporate the instructional and assessment techniques directly into the program materials and procedures, thereby increasing the probability of high learner achievement of the instructional objectives.

The IS is primarily useful in specifying instruction prior to the development of materials and procedures. However, the structure and architecture of extant instruction and curricula are seldom explicitly stated. Postdictive analytic conventions (Smith, 1972) have been developed for use in analyzing the instructional architecture of portions of instructional materials. Set and matrix notational conventions permit description of extant material in terms of the following seven components:

Elements: the phenomena to be described, compared, related, or otherwise studied (e.g., objects, systems, events, groups).

Variables: the characteristics of properties of elements that are used to describe, compare, and relate them (e.g., color, weight, cost).

Values: the terms, phrases, numbers, or other symbols which are available for assignment to elements for a given variable (e.g., red, 4 pounds, 50¢).

Describers: those values of variables which are assigned to particular elements.

Observation/Measurement Procedures: standard procedures or algorithms used to assign values of variables to particular elements (e.g., using a thermometer to measure the temperature of a liquid).

Relational Rule: rules or algorithms which specify describers for one variable given describers for another variable (e.g., $A = \pi r^2$, all the rectangular blocks are green).

Correspondence Rules: sets of rules used to relate one set of elements to another set of elements (e.g., the letter *p* is pronounced /p/).

Text-Referenced Instructional Management Systems

Tests and texts have traditionally been treated as independent units with given tests amenable to various texts and the outcomes of instruction with a given text assessable by various tests. It is possible, however, to produce tests referenced to a given text series. With the test directly coupled to the text a means is provided for determining the extent to which specific outcomes are being attained by individual students after specified instruction. It is also possible to prepare supplementary practice materials referenced to each criterion measure for use where adequate proficiency is yet to be attained. This integrated sequence of "text-test-troubleshooting materials" constitutes a simple instructional management system, which SWRL for convenience has termed a Learning Mastery System (LMS).

A prime limitation in producing such systems is that current texts rarely have clear statements of instructional outcomes. This limitation has been met by inferring the measurable outcomes associated with a given text. Although simple in structure and use, and LMS significantly expands the information available to the teacher for instructional decisions. Each LMS provides:

A means for student placement at the beginning of the school year

Criterion-referenced measures on three to eight instructional outcomes ten to fifteen times during the year

Additional practice materials for the outcomes which have continuity throughout the text

Mid-year and end-of-year evaluation measures

Multiple Matrix Sampling

The specific equations used in multiple matrix sampling provided by Lord (1960) and Lord and Novick (1968) have been procedurally adapted for implementation (Shoemaker, 1973) and applied to large-scale group achievement assessment. Results to date indicate that parameters estimated through multiple matrix sampling and parameters obtained through testing all examinees on all items may be interpreted similarly. Parameters estimated through multiple matrix sampling may be contrasted with any predetermined standard defining the minimal level of acceptable achievement.

State-of-the-Art Statistical Analyses

SWRL's research and development activities require on-line access to large data files and considerable flexibility in manipulating, analyzing, and retrieving information. In addition to standard statistical and matrix manipulation utility packages, a capability has been developed for the continuous upgrading of an extensive library of computer program-building modules. This permits quick modification of computer program functions with a minimum of reprogramming for new procedures defined by staff.

MATTERS THAT ARE ON THE LEADING EDGE OF SWRL STATE-OF-THE-ART

The areas and activities described in this section include items that, while not quite available as "shelf-items," will influence the "new" generation of SWRL instructional products.

Quality Assurance Systems

Before a SWRL-developed instructional program is released it must be demonstrated that it has been used successfully to obtain prespecified levels of pupil performance. To provide a replicable means of ensuring that the program continues to function at these levels, a set of procedures referred to as Quality Assurance—QA— (Hanson, 1972) has been developed. These procedures provide enroute information on various indicators of performance and pacing useful to teachers, principals, and district administrators. Teachers have benefited from QA because it provides information helpful in planning and pacing instructional activities throughout the school year. Principals and district personnel find Quality Assurance helps keep them informed of the status of an instructional program in each class throughout the school year. Pupils also benefit because it provides teachers with the assistance needed to complete all instructional units and to achieve high performance on the major outcomes.

Integrated Instructional Information Systems

Text-referenced instructional management systems assign the teacher total accountability for the attainment of instructional outcomes. While the teacher often passes the responsibility on to the students, and occasionally to other school personnel, the teacher at present is the sole manager of instruction.

The confounding of the teacher, instructional materials, and instructional decisions in assessing accountability fails to recognize that the teacher shares responsibility for the instructional progress of students with administrators at the school and district level, and with parents. It is possible to provide useful information to each of these groups. However, the mechanisms for such information provision are sufficiently complex to require automation of analysis and reporting functions. This is the scope of the SWRL Instructional Management System (IMS).

The SWRL Instructional Management System operates in conjunction with a developed instructional system such as the SWRL/Ginn Kindergarten Program or with an application of the Text-Referenced Management System. Utilizing a variety of communication modes for input and output, reports for each category of individuals requiring information are specially designed to aggregate and synthesize the information in a manner that is understandable and comprehensive, and consistent with the need-to-know requirements of teachers, principals, curriculum supervisors, district administrators, parents, students, and development personnel (McManus, 1972).

Program-Fair Evaluation

In the SWRL context "program-fair" simply indicates that all assessment procedures are systematically referenced to the particular objectives of the program and the stimulus content used in instruction are related to the objectives. Shoemaker (1972) has reviewed the state-of-the-art in this area. These techniques provide fair approximations for "program-fair" comparisons of instructional programs. The adequacy of the approximations can only be assessed after the techniques have been further exercised empirically.

The Architecture of Instructional Programs

The item form and the instructional specifications (IS) are useful tools in instructional product development, but they are neither necessary nor sufficient to initiate or advance a given product development effort. Sets of IS's make it possible to define "trees" at an intermediate level of complexity above the micro-level of behavioral objective "twigs," but below the macro-level of an architectural framework. The architectural framework of an instructional program converts the "jungle" of instruction into an orderly "forest" configuration.

Emulating established procedures in the architecture of physical structures, the architecture of instruction can be conducted in stages of schematic specifications, through preliminary specifications to working specifications. Instructional architecture subsumes the planning of "skills" and "content" conventionally considered in test design. Statements of instructional architecture are as yet few and far between. Examples of preliminary specifications can be found in Quellmalz (1972). An example of working specifications can be found in a SWRL (1972) document prepared by Baker, drawing upon some previous SWRL papers.

Instructional Development Control and Monitoring System (IDCMS)

IDCMS represents an integrated hardware configuration presently being installed within the Laboratory facility. It represents a powerful tool for increasing the sophistication of educational research and development activities. Computer applications to behavior research have typically been restricted to statistical analyses of data collected off-line. This type of requirement can be handled by standard statistical and matrix manipulation utility packages. Although such a capability is important, Laboratory product design requirements include studies of real time interactions between subject and equipment. Exploitation of IDCMS capabilities will permit online experimentation involving complex event sequences, variable media utilization, and real time test contingencies. Figure 1 includes a block diagram of the IDCMS configuration.

MATTERS THAT ARE BEYOND SWRL STATE-OF-THE-ART

In the context of research and development an instructional product that completes all stages of the development cycle is considered final; the "now" generation of the product unashamedly represents the best that can presently be delivered. However, long before the "now" product has gone to market the outcomes of programmatic research and development activity provide the scientific and technological bases for the "new" generation. Listed below are some items that, were they now even "leading edge," the description of "new" generation products would likely be dramatically different. Yet until they are classed as available shelf items the "new" generation of instructional products cannot be expected to reflect them.

1. Instructional data base structures in fields other than mathematics and reading.

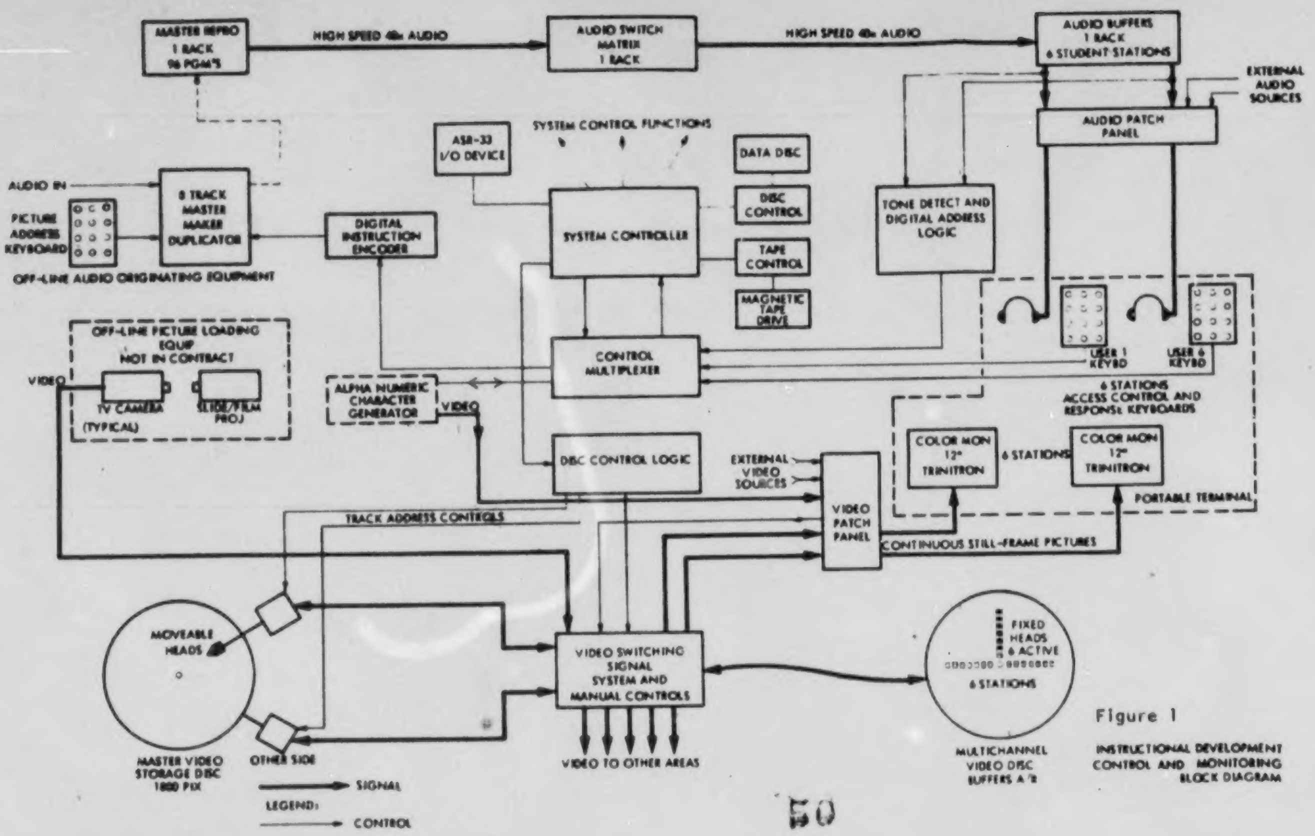

MASTER REPRO
1 RACK
96 PGM'S
HIGH SPEED 40x AUDIO
AUDIO SWITCH
MATRIX
1 RACK
HIGH SPEED 40x AUDIO
AUDIO BUFFERS
1 RACK
6 STUDENT STATIONS
EXTERNAL
AUDIO
SOURCES
AUDIO PATCH
PANEL
SYSTEM CONTROL FUNCTIONS
ASR-33
I/O DEVICE
DATA DISC
DISC
CONTROL
TAPE
CONTROL
MAGNETIC
TAPE
DRIVE
SYSTEM CONTROLLER
TONE DETECT AND
DIGITAL ADDRESS
LOGIC
AUDIO IN
PICTURE
ADDRESS
KEYBOARD
8 TRACK
MASTER
MAKER
DUPLICATOR
DIGITAL
INSTRUCTION
ENCODER
OFF-LINE AUDIO ORIGINATING EQUIPMENT
OFF-LINE PICTURE LOADING
EQUIP
NOT IN CONTRACT
VIDEO
TV CAMERA
(TYPICAL)
SLIDE/FILM
PROJ
ALPHA NUMERIC
CHARACTER
GENERATOR
VIDEO
CONTROL
MULTIPLEXER
USER 1
KEYBD
USER 6
KEYBD
6 STATIONS
ACCESS CONTROL AND
RESPONSE KEYBOARDS
DISC CONTROL LOGIC
EXTERNAL
VIDEO
SOURCES
VIDEO
PATCH
PANEL
COLOR MON
12"
TRINITRON
6 STATIONS
COLOR MON
12"
TRINITRON
PORTABLE TERMINAL
TRACK ADDRESS CONTROLS
CONTINUOUS STILL-FRAME PICTURES
MOVEABLE
HEADS
VIDEO SWITCHING
SIGNAL
SYSTEM AND
MANUAL CONTROLS
FIXED
HEADS
6 ACTIVE
6 STATIONS
MASTER VIDEO
STORAGE DISC
1800 PIX
OTHER SIDE
VIDEO TO OTHER AREAS
MULTICHANNEL
VIDEO DISC
BUFFERS A/B
SIGNAL
LEGEND:
CONTROL

Figure 1

INSTRUCTIONAL DEVELOPMENT CONTROL AND MONITORING BLOCK DIAGRAM

2. Systemic structures in the social domain.
3. Cost-feasible automated interactive instructional hardware/software systems.
4. Algorithms for prespecified instructional decision contingencies.
5. Quality control systems for aspects of performance other than qualitative attainment, time, and cost.

That the development of these items will involve measurement considerations is clear. These considerations move far from such classical topics as validity, reliability, item analysis, norming, and other traditional tools of psychometric theory and practice. It is well to have these tools in the instructional development kit, but more sophisticated tools are clearly needed.

REFERENCES

Baker, R.L., & Schutz, R.E. (Eds.) *Instructional product development.* New York: Van Nostrand Reinhold, 1971.

Bruner, J.S. *The process of education.* Cambridge, Mass.: Harvard University Press, 1960.

Ebel, R.L. Criterion-referenced measurements: Limitations. *School Review,* 1971, *79,* 282–288.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist,* 1963, *18,* 519–521.

Hanson, R.A. The contribution of quality assurance procedures to laboratory and user evaluation of educational programs. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1972.

Hively, W. *Defining criterion behavior for programmed instruction in elementary mathematics.* Cambridge, Mass.: Harvard University, Committee on Programmed Instruction, 1963.

Lennon, R.T. Assumptions underlying the use of content validity. *Educational and Psychological Measurement,* 1956, *10,* 294–304.

Lord, F.M. Use of the true-score theory to predict moments of univariate and bivariate observed-score distributions. *Psychometrika,* 1960, *25,* 325–342.

Lord, F.M., & Novick, M.R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

McManus, J. *Technical note: Scope of IMS-3.* SWRL Educational Research and Development, Los Alamitos, Calif., 1972.

Quellmalz, E. Structure of a fine arts instructional program. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February, 1973.

SWRL. *Technical specifications: English language and concepts for Spanish speaking children.* SWRL Educational Research and Development, Los Alamitos, Calif., 1972.

Schutz, R.E. Criterion-referenced testing. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1972.

Shoemaker, D.M. Evaluating the effectiveness of competing instructional programs. *Educational Researcher,* 1972, *1,* 5–8.

Shoemaker, D.M. *Principles and procedures of multiple matrix sampling.* Cambridge, Mass.: Ballinger, 1973.

Smith, E.L. Procedures for generating candidates for learning hierarchies. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Chicago, April, 1972.

Sullivan, H.J., et al. Developing instructional specifications. In R.L. Baker & R.E. Schutz (Eds.), *Instructional product development.* New York: Van Nostrand Reinhold, 1971.

GENERATING CRITERION-REFERENCED TESTS FROM OBJECTIVES-BASED ASSESSMENT SYSTEMS: UNSOLVED PROBLEMS IN TEST DEVELOPMENT, ASSEMBLY, AND INTERPRETATION

Rodney W. Skager
Center for the Study of Evaluation

We begin with the notion that criterion-referenced tests, whatever their other characteristics, are by definition "objectives-based." This is clearly implied in Glaser and Cox (1968) and elsewhere. Compared to the content forming the traditional test plan, performance objectives delineate domains of test content that are considerably less open to differing interpretations by those who develop assessment materials. Performance objectives also tend to legitimate test items that in the more traditional usage might be rejected due to extreme difficulty values for a given population of examinees. In this regard, Cox (1965) has shown that the way in which item-difficulty indices are typically used can result in significant differences between actual and intended test content. Performance objectives, furthermore, become the primary means by which scores on criterion-referenced tests are interpreted.

Performance objectives have two essential characteristics. First, they contain an action statement incorporating a verb describing some sort of observable behavioral output. Second, they incorporate a description of the conditions and materials with which the examinee is to perform the action. The following objective from the System for Objectives-Based Assessment—Reading collection (SOBAR Field Manual I, 1972) illustrates:[1]

> After listening to the story, the learner will sequence illustrations of the main events of the story. (41204)

The above discussion goes over rather familiar ground. Still, the knowledge that an item pool is derived from an explicit performance statement may generate a false sense of orderliness. It turns out that considerable heterogeneity may exist within the item pool, particularly with respect to item difficulties for a given population.

If one were to examine the variety of performance objectives being

---

Preparation of this paper was supported by funds from the National Institute of Education (NIE), Department of Health, Education, and Welfare. Points of view or opinions expressed do not necessarily represent official NIE position or policy.

[1] Many writers suggest that behavioral or performance objectives should also contain a statement about the standard of performance to be achieved by the examinee, e.g., "answer 8 out of 10 items correctly." The standard of achievement appears to be an entirely different issue than the description of the performance itself. Indeed, the question of how such standards are to be established obviously remains a major unresolved issue.

developed around the country it would quickly be apparent that various collections differ sharply in the specificity of the performance domains defined by individual objectives. It is quite possible to write two objectives of excellent quality, each with explicit conditions and action statements, yet one of the objectives defines a very limited content domain—perhaps only a single assessment item—while the other defines a virtually infinite population of test items quite heterogeneous in difficulty for a given group of learners. Such a likelihood was anticipated by Krathwohl (1965) who described objectives at several levels of generality and even earlier by Gagné (1963) in his distinction between "mastery" and "transfer" objectives.

ASSESSMENT AND INFORMATION NEEDS

A second notion of this paper is that those who construct achievement tests tend to build one kind of instrument without regard for the variety of uses to which that instrument might be put. In an earlier paper (Skager, 1972) I discussed six major information needs in education which depend upon instruments designed to assess pupil progress. These are:

1. *Planning the curriculum:* What is the content of the tests which will later be used to assess the effectiveness of instruction?
2. *Classroom Management:* What is the present learning status of the pupils in terms of the objectives and prerequisites of the instruction?
3. *Evaluating Instruction:* What is the terminal learning status of students who have been exposed to the program or one or more of its sub-units?
4. *Accountability:* What is the terminal learning status of students who have been instructed by particular teachers or in particular schools?
5. *Allocating Resources:* Where are deficiencies in the achievement of students so severe as to require the allocation of additional effort and funds?
6. *Prediction:* What will be the future achievement of individuals in particular educational or employment situations?

The earlier paper (Skager, 1972) argued that contemporary published tests do not do a good job of meeting several of the above information needs, but not because today's tests are poorly constructed. A single type of instrument, especially one developed primarily for (6) above, simply cannot meet all of the information needs cited. Tests built for the purpose of providing feedback to the teacher in the classroom would undoubtedly be of the criterion-referenced type. Each test would probably measure only one or at most a few related instructional objectives. Instruments built for large-scale resource allocation studies would in most cases measure many objectives and consist of a number of separate test forms derived through item-sampling procedures. Such instruments might be referenced to performance criteria as well as to norms.

OBJECTIVES-BASED ASSESSMENT SYSTEMS

A third notion of this paper, one that follows from the observations in the previous section, is that questions relating to the construction and interpretation of criterion-referenced tests ought to be formulated in light of a new model of test development and assembly. There appears to be an unfortunate tendency on the part of test publishers to view criterion-referenced tests as just another kind of published test. That is, tests are being prepared and printed for mass sale with the idea that the same set of items will be used over and over again and without regard for the type of information needed from the assessment, in just the same way that contemporary norm-referenced instruments are used. This appears to me to be a seriously outmoded conception.

Four years ago Wood and Skurnik (1969) published an important paper on the potential of item banking as a tool for generating tests for particular purposes. The item-banking concept offers great flexibility when it is incorporated into a system for generating tests with desired content and of appropriate length for various information needs. Such a system could rapidly identify a desired number of items measuring a single performance objective or a specified combination of objectives. Large-scale assessment projects could utilize item-sampling procedures, producing information on the status of learners with respect to many objectives, yet requiring only a minimum of assessment time per student. Instead of the single test used thousands or even millions of times, the item bank has the potential for generating an infinite number of different tests, each tailored to meet a particular set of information needs.

The System for Objectives-Based Assessment—Reading (SOBAR),[2] under development at the Center for the Study of Evaluation, is basically an item bank integrated into a selection/delivery system. Described elsewhere, (see, for example, Skager, Borgerding, & Dahl, in press) SOBAR consists of three basic components:

1. A bank of assessment items, each a member of a sub-pool keyed to a performance objective.
2. A set of performance objectives designed to cover the entire spectrum of a content domain, in this case that of reading, grades K through 12.
3. A classification system used for selecting particular sets of objectives and for score reporting at various levels of generality.

---

[2] Since this paper was written a contract has been signed with Science Research Associates, Inc. (a subsidiary of IBM), to publish SOBAR. SRA is doing additional development work on the delivery of the system, scoring, reporting, etc.

To establish an operational system, other types of components are also needed. These include, for example, user's manuals which describe strategies for selecting objectives and for specifying the type of assessment instruments that are needed in light of the information needs. Also useful are computer-based item retrieval systems as well as scoring and reporting services. Taken together, these components represent a system for generating criterion-referenced tests designed for different purposes, in sharp contrast to today's "testing programs" built around a single published instrument.[3]

CRITICAL ISSUES

The preceding discussion provides a context for identifying critical technical issues likely to be encountered in building and interpreting criterion-referenced tests. Three considerations were advanced. (1) Criterion-referenced tests are properly objectives-based, but in many cases this still allows for considerable heterogeneity in the difficulty of the items developed to measure any given objective; (2) If criterion-referenced tests are designed specifically to meet particular information needs in education, then tests developed for various needs will differ radically in number of items needed to measure an objective as well as in number of objectives covered by the test; (3) The concept of an objectives-based assessment system was advanced as a means for generating criterion-referenced tests designed for particular information needs in education.

The implications of all three of these notions for criterion-referenced testing is that the field will (or at least should) get away from the notion of a single test used (or misused) for a variety of information needs, and instead develop flexible, multi-purpose test-generation systems utilizing item banks referenced to sets of explicit performance objectives. This last sentence defines what I view as the larger issue. If we do not orient ourselves to that larger issue there is the danger that we will fail to recognize issues that did not arise under the old model of test development.

UNRESOLVED QUESTIONS

The question is not simply how criterion-referenced tests are to be developed and interpreted, but rather how objectives-based assessment systems, utilizing item banks, may be used to develop such instruments in forms that will meet various information needs in education.

Six questions are discussed below. Each has already been raised in relation to CSE's SOBAR project, some by staff and others by interested outsiders. Interestingly, it turns out that each question has its analogue in contemporary test theory and practice.

---

[3] There is no reason why an objectives-based system such as SOBAR could not be used to generate tests which could also be interpreted in the norm-referenced sense. A means for specifying desired test content plus normative data on the items in the bank, however, would be needed.

1. *Independence:* Given the rationally derived structure of a content domain, a set of performance objectives devolving from that structure, and pools of assessment materials written to measure each objective, is there any need to verify empirically whether or not the performances specified by the objectives are sufficiently independent from one another to provide non-redundant information?

The question is not concerned with the empirical validation of a rationally conceived learning hierarchy such as that proposed by Resnick and Wang (1969). The performance objectives referred to here represent terminal points of instruction, not the sequence of steps taken to get there. Here we have a familiar question about the factor structure of a set of assessment items, but it is posed on a rather grand scale in the case of systems incorporating objectives covering entire content domains such as reading or mathematics as taught in the elementary and secondary schools. And it is not simply a matter of one's intellectual preference for the state of grace implied by factor purity. Assessment is expensive in terms of time and money. Unnecessary assessment generating redundant information ought to be avoided.

Dahl (1971) selected a subset of eight potentially related objectives from the initial SOBAR collection. Brief item pools measuring each objective were administered to children. The resulting item intercorrelations were analyzed by two separate factor analytic procedures and the items for two of the objectives clearly loaded on the same factor in both analyses. On inspection it was observed that the objectives indeed called for similar types of performance on somewhat related content. In one case the child was asked to generate words beginning with certain consonants. In the other case the child was given consonant digraphs and asked to perform the same task.

Judges were also given the items and objectives separately and asked to classify each item under the objective it was written to measure. The judges had no trouble performing this task and even managed to distinguish between the two item pools which were indistinguishable in the factor analysis of examinee responses. Unfortunately, item pools measuring one or two hundred objectives which might be in the curriculum at a given age level exceed our data processing capabilities as well as children's capacity to take tests. Still, rational analyses of complex domains of content are inevitably arbitrary, even if they seem sensible to their creators and are backed by the opinion of other experts. An independent analysis of the same content domain would certainly not come up with an identical classification system or set of objectives.

2. *Validity:* How does one establish the fact that the items in the pool measuring any objective are valid in the sense of being (a) congruent with the objective, e.g., actually measuring the performance described

in the objective, and (b) comprehensive in the sense of providing adequate coverage of the domain specified by the objective?

This question and the preceding one are obviously related. A perhaps artificial distinction has been made in the case of the congruence issue on the reasoning that it is possible to have two objectives which are reasonably independent, and yet find that at least some items nominally classified under these objectives do not share a sufficiently large proportion of common variance with other items classified under the same objective.

The problems of objective-item congruency and comprehensiveness are significant concerns in the development of objectives-based item pools. It is one thing to maintain that one's item writers are competent and careful; it is another to prove it empirically. Obviously, there are already a variety of systematic techniques for getting at the matter of *congruence*. In the study mentioned above Dahl (1971) found items in a number of instances which correlated more highly with items written to measure other objectives than with other items in their own pool. Guttman and Schlesinger (1966) utilized a form of cluster analysis which produced "coefficients of similarity" between pairs of items. This easily interpreted statistic can be used to construct plots revealing the extent to which anticipated patterns actually occur. There are obviously many approaches to assessing the degree of common variance shared by items measuring a given objective. The question of which methods are the most appropriate needs to be addressed.

Establishing the *comprehensiveness* of the items written to assess the domain defined by an objective would presumably be approached through judgmental procedures. Of course, independent performance measures might be devised based on the notion that a person who answered all of the items in the domain correctly ought to be able to manifest the skill defined by the objective in some other way as well. Still, it is not entirely clear in a given case whether a positive result would reflect the comprehensiveness of the item pool itself, or instead reflect the predictive validity of the existing items in that pool. The potential costs of this approach, especially for complex content domains, are staggering.

3. *Identifying "Bad" Items:* How does one identify poorly written items by means of item analysis procedures when the frequency of correct response may be extremely high or low, accurately reflecting the achievement status of a particular group of learners?

It has been recognized for some time that deviant values on item difficulty and discrimination indices are insufficient conditions for concluding that a criterion-referenced item is "bad." Unfortunately, items that are poorly written in terms of the criteria listed in typical measurement text-books can still occur. Even though equipped with the explicit guide a clearly written performance objective provides, item writers will still now and then write misleading instructions, include more than one correct alternative, inadvertently provide unintended clues as to which is the correct answer, and the like. When developing assessment materials for children there is the additional problem brought on by the extensive use of pictures and drawings. Does the picture represent the same thing to the child that it does to the adult? (e.g., will the child see the object as a table rather than a bench?) If the name of some object in the picture is important, one had better be certain that there is only one word in the child's vocabulary for that object. (On the SOBAR project we have had to discard a considerable number of illustrations because a "sofa" is also a "couch," etc.)

Relying solely upon judgments as an index of item quality ought to leave us just as uneasy in the case of criterion-referenced tests as it should for norm-referenced instruments. Of course, one solution might be in the suggestion that the test developer can always find some group with the relevant knowledge or skill only partially developed. If it could be assumed that the proportion passing each item should lie somewhere approximately in the range $p = .3$ to $.7$, traditional item statistics would be applicable. In the case of many useful collections of objectives, this situation would be improbable. Suppose one were assessing the quality of a pool of items for an objective involving decoding initial consonants when followed by certain vowels. One could conceivably find a group of learners for whom the average proportion passing each item was .5. The problem is that the average item difficulty might well be a reflection of the fact that the learners have virtually mastered some consonants while they still have not achieved others. This situation, not at all unlikely in the real world of the classroom, would produce a very mixed bag of item statistics.

A number of authors have attempted to get around the problem of low quality items and items which are not congruent with the objectives they are supposed to measure by developing systematic item-generation procedures which eliminate the role of error-prone human item writers. Item-generation rules, such as those used by Guttman and Schlesinger (1967), may offer a way of developing tests not requiring empirical verification of congruency and item quality. The general applicability of such procedures for the majority of instructional content domains still seems open to debate, however. Some sort of resolution of this issue with respect to the item-generation procedures presently available would be extremely helpful.

4. *Information on Items in Bank:* Assuming that the items in a bank have met necessary tests of quality, what sort of information might be stored on each that would aid in constructing tests and interpreting the scores which would eventually result?

The above question really does not represent so much a "problem" as it does a potential. Item banks integrated into objectives-based assessment systems will offer a variety of options among different strategies for assembling tests. The basic rule for criterion-referenced tests is that they be made up of items which measure some pre-determined set of objectives. But there are other interesting rationales for item selection. For example, one might want a test assessing objectives that are attained by some reference group at a given age or grade level. Such an instrument would be amenable to both criterion- and norm-referenced interpretations. One might conceivably want to construct a test which could be used to predict performance on a *different* set of objectives, perhaps ordinarily attained at some point later in time. Or one might want to use existing item data to build a strictly normative instrument without any particular concern about interpretation of the results in terms of what instructional content has been attained.

The question, then, is really one of anticipating what types of item data might be collected as a part of the regular operation of an objectives-based assessment system or in specially designed research studies. This determination would stem from an analysis of the various kinds of interpretations of test scores one might conceivably want to make *in addition to* the familiar one of attainment vs. non-attainment of an objective. Given the right kinds of item data, a computer could be programmed to select items maximally appropriate for any of the desired modes of interpretation. The possibility of describing items by means of parameters which are invariant with respect to the group of learners on whom the data were derived is naturally appealing, as was recognized earlier by Wood and Skurnik (1969).

5. *Sequencing Objectives:* When the collection of objectives represents terminal points in instruction is it necessary and appropriate to find some "ideal" sequence by which instruction might proceed?

A multipurpose objectives-based assessment system is not a design for instruction. Such a system does, however, define endpoints of potentially separate units of instruction. Concentrating on objectives of a terminal type, it thus does not incorporate sequenced "enabling objectives" developed for a particular instructional program. As a result, the need for precise sequencing is not so apparent as it would be for specific curriculum hierarchies like those Gagné (1962) has described. In spite of this, field experience with SOBAR materials has repeatedly shown that people involved in instruction usually request recommendations on the sequencing of the objectives they have selected.

There is a certain intuitive sequence apparent in any major content domain incorporating hundreds of objectives. In reading, one would deal with encoding and decoding before dealing with paragraph meaning. But within more specific categories there is not necessarily an obvious order, and one suspects that alternative sequences might be equally appropriate, and that instruction on several objectives might proceed at the same time. In most classrooms one would expect that objectives from different major categories would be taught simultaneously.

If the decision were made to define an appropriate sequence (or sequences) of objectives one would be faced with the usual choice between depending on expert opinion or seeking empirical verification on learner performance. The latter would involve a formidable expenditure of time and resources, especially if alternative sequences were to be compared. One cannot help but wonder if it is even feasible for comprehensive, "program-independent" systems like SOBAR. Nevertheless, if recommended sequences are not provided, it is inevitable that sequencing will be done in local schools and school districts, ordinarily without credible attempts at verification.

6. *Defining Mastery:* How many items does one include on criterion-referenced tests when the purpose is to determine whether learners have achieved mastery of an objective (or objectives), taking into account (a) the generality of the item pool in terms of the variety of performances defined by the objective, (b) whether the response called for is to *produce* the right answer or *select* the right answer, and (c) whether the resulting information will refer to individual learners or groups of learners?

How to define the nature of any performance that would indicate mastery of a domain of content remains a major conceptual problem. To arbitrarily rule that mastery is indicated when a certain number of items is passed on a test merely avoids the larger issue. From the point of view of a learning psychologist mastery might well mean that the learner is now ready to learn or do something else. That is, one can predict that transfer will occur.

A discussion of the validity issue by Harris later in the monograph points out that information about transfer effects within learning hierarchies is relevant only in the case of assessment devices linked to particular instructional systems. With program-independent systems like SOBAR it would not be feasible to define mastery of a given objective (or unit) in terms of the probability of success on the next objective (or unit). The deliberate comprehensiveness of the collection of objectives means that all would not be selected for a given instructional program and that, as suggested above, those that are selected could be sequenced in a variety of ways.

Given a definition of mastery that is both conceptually and technically acceptable, it is still necessary to decide how many items should be included on a given test. The specificity/generality of the objective(s) on which a criterion-referenced test is based must be considered when items are assembled and scores interpreted. If the actual item pool is very large, or potentially very large, and if items in that pool differ in difficulty for a group of learners, then we would presumably need more items on the test than would be necessary for a very small item pool with individual items very similar in difficulty for any given group of learners. In traditional terms, we are concerned about *reliability* in the sense of consistency or equivalence. If an insufficient number of items is included on the test, then another sample of items taken from the same pool might give a markedly different result. This approach to the concept of reliability appears to me to have the most pragmatic relevance to objectives-based evaluation systems incorporating item banks. However mastery itself is defined, the question remains, "How many items (i) should be included on the test so as to assure within certain limits that the learner (or learners) would obtain the same score on another sample of *i* items from the same pool?" There is obviously nothing new in defining reliability in terms of the relationship between randomly parallel measures administered with zero time interval. But this empirical approach to the concept seems to focus on the most relevant information for the type of assessment system being considered here.

If the test incorporates production type items, where guessing is an improbable or even impossible basis for a correct response, fewer items are required. Yet if one looks at the literature which has begun to accumulate around the topic it is quite apparent that the model of the traditional multiple-choice test item is the only one being considered. This seems odd, since a valid analysis of virtually any major school content domain will produce many performance objectives which require the learner to generate a response rather than to select one from among a set of alternatives.

The type of decision for which the assessment information is to be used is possibly the most important factor of all, whether or not one is making that decision about individuals, small groups, or very large groups. At present a good deal of attention is being devoted to the single question of how many multiple-choice items should be included on a test to assess a pre-determined, but theoretically unsubstantiated, criterion of mastery. For example, Millman (1972) used the binomial model to generate tables for determining appropriate passing scores for tests of different lengths without regard to the generality of the domain, whether or not the items are selection or production, and the type of decision to be made from the resulting score.

The work by Millman and others is obviously meant to facilitate the one type of decision—that made by the teacher about an individual child's acquisition of an instructional objective. It would of course be useful to have this question answered, especially in terms of the parameters suggested above. Glaser and Nitko (1971) discuss the possible application of acceptance sampling and sequential likelihood-ratio tests, methods which might well be useful, even essential, in computer-assisted instruction, but which are hardly likely to be regarded with enthusiasm by the classroom teacher. The relative "risk" involved in passing a learner who in reality has not quite mastered an objective according to an arbitrary criterion of performance, or of continuing instruction on an objective that has in reality been mastered, does not seem unduly severe given that the teacher is equipped with a reasonable degree of professional judgment. This argument will obviously not appeal to many readers with a psychometric orientation.

Much more critical is the decision about how many items to include in the test in the first place. In the conception of an objectives-based assessment system presented in this paper that decision is not made in the field. Rather, it is a dilemma facing those who assemble tests from the item bank. At this point in the development of systems like SOBAR it is a serious dilemma indeed.

SUMMARY

This paper has focused on the idea that technical problems relating to the criterion-referenced test ought to be defined in terms of the larger issues of (a) the purposes for which tests are used in education as well as (b) the nature of assessment systems which may in the future be used to generate such instruments. Specifically, it has been suggested that contemporary standardized testing programs do not adequately meet several important information needs in education. The concept of the objectives-based assessment system, incorporating performance objectives, item banks, information storage, and test assembly capabilities has been advanced as a modern alternative. Six technical problems or issues associated with criterion-referenced tests generated from such systems have been discussed.

REFERENCES

Cox, R.C. Item selection techniques and evaluation of instructional objectives. *Journal of Educational Measurement,* 1965, 2, 181–185.

Dahl, T.A. The measurement of congruence between learning objectives and test items. Unpublished doctoral dissertation, University of California, Los Angeles, 1971.

Gagné, R.M. The analysis of instructional objectives. Paper presented for the National Symposium on Research in Programmed Instruction, Department of Audiovisual Instruction, National Education Association, 1963.

Gagné, R.M., Mayor, J.R., Garstens, H.L., & Paradise, N.E. Factors in acquiring knowledge of a mathematical task. *Psychological Monographs,* 1962, *76* (Whole No. 526).

Glaser, R., & Cox, R.C. Criterion-referenced testing for the measurement of educational outcomes. In R.A. Weisgerber (Ed.), *Instructional process and media innovation*. Chicago: Rand McNally, 1968.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Guttman, L., & Schlesinger, I.M. Development of Diagnostic, Analytical and Mechanical Ability Tests Through Facet Design and Analysis. U.S. Office of Health, Education and Welfare, Project No. OE-15-1-64, 1966.

Guttman, L., & Schlesinger, I.M. Systematic construction of distractors for ability and achievement testing. *Educational and Psychological Measurement*, 1967, *27*, 569–580.

Krathwohl, D. Stating objectives appropriately for program, for curriculum, and for instructional materials development. *Journal of Teacher Education*, 1965, *16*, 83–92.

Millman, J. Passing scores and test lengths for domain-referenced measures. Los Angeles: Instructional Objectives Exchange, 1972.

Resnic, L.B., & Wang, M.C. Approaches to the validation of learning hierarchies. Paper presented at the Eighteenth Annual Western Regional Conference on Testing Problems, San Francisco, May, 1969.

Skager, R.W. Information gaps in education: Objectives-based evaluation systems as alternatives to today's testing programs. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Skager, R.W., Borgerding, J., & Dahl, T.A. The system for objectives-based assessment reading: Rationale, components, and uses. Los Angeles: Center for the Study of Evaluation, University of California. (In press)

SOBAR field manual I. Prepared by the staff of the Program for Research on Objectives-Based Evaluation. Los Angeles: Center for the Study of Evaluation, University of California, 1972. (Field Test Edition)

Wood, R., & Skurnik, L.S. Item banking. Slough, England: National Foundation for Educational Research, 1969.

PROBLEMS IN THE DEVELOPMENT OF CRITERION-REFERENCED TESTS: THE IPI PITTSBURGH EXPERIENCE

Anthony J. Nitko
Learning Research and Development Center
University of Pittsburgh

Criterion-referenced testing has come about on a new wave of psychology—a psychology expressing an increasing concern for instruction and the instructional process. Such an instructional psychology postulates a theory of instruction that is prescriptive with respect to the instructional procedure itself. A learning theory, on the other hand, is descriptive and after the fact, specifying the conditions under which the learning occurred (Burner, 1966).

> In theories of instructional psychology primary focus is on . . . (a) a description of the state of knowledge to be achieved; (b) description of the initial state with which one (i.e., the learner) begins; (c) actions which can be taken, or conditions that can be implemented to transform the initial state; (d) assessment of the transformation of the state that results from each action; and (e) evaluation of the statement of the terminal state desired (Glaser & Resnick, 1972, p. 208).

Glaser's motive for applying criterion-referenced testing to educational achievement measurement (Glaser, 1963) stemmed from a concern about the kind of achievement information needed to make instructional decisions from the above kind of instructional psychology. Some instructional decisions concern individuals and may relate, for example, to the kind of competence an individual needs in order for him to be successful in the next course of a sequence. Other decisions center around the adequacy of the instructional procedure itself. Tests that provide achievement information about an individual only in terms of how the individual compared with other members of the group tested, or which provide only sketchy information about the degree of competence the individual possesses with respect to some desired educational outcome, are not sufficient to make the kinds of decisions necessary for effective instructional design and guidance.

Glaser's (1963) application combined both the notion of a desired model

---

Preparation of this paper was supported by the Learning Research and Development Center, which is supported in part by funds from the National Institute of Education (NIE), U.S. Department of Health, Education and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred. Grateful acknowledgement is made to Drs. Cooley, Cox, Glaser, Hsu, and Resnick for their helpful comments on the draft manuscript.

or a minimum goal we would like an individual to attain (Flanagan, 1951) and the notion of a standard domain of content (Ebel, 1962). He called for the specification of the type of *behavior* the individual is required to demonstrate with respect to the content. This distinction between behavior or performance and content is at the heart of criterion-referenced testing. "The standard [or criterion] against which a student's performance is compared . . . is the behavior which defines each point along the achievement continuum (Glaser, 1963, p.519)." *A criterion-referenced test, then, is one that is deliberately constructed to give scores that tell what kinds of behavior individuals with those scores can demonstrate* (Glaser & Nitko, 1971).

Note that this definition does not imply a predetermined, fixed cutting-score (cf. e.g., Livingston, 1972); it does not imply simply writing a set of behavioral objectives and keying a set of items to those objectives; and it does not imply the use of only open-ended production items (cf. Harris & Stewart, 1971). The definition, instead, implies that there are four characteristics inherent in criterion-referenced tests:

> The classes of behaviors that define different achievement levels are specified as clearly as possible before the test is constructed.
>
> Each behavior class is defined by a set of test situations (that is, test tasks) in which the behaviors can be displayed in terms of all their important nuances.
>
> Given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test.
>
> The obtained score must be capable of expressing objectively and meaningfully the individual's performance characteristics in these classes of behavior (Nitko, 1970).

These four characteristics form the central theme of this paper. The focus is on the development of criterion-referenced tests having these properties and some associated technical problems that are encountered. Solutions for these technical problems are not readily available nor immediately generalizable to all curricular areas for which criterion-referenced tests might be desired. Attempts are made, therefore, to specify procedures that will be useful to the practical developer until the technical problems are solved.

The characteristics outlined above appear to form a logical developmental sequence. This sequence is seldom followed in practice. In fact, a great deal of criterion-referenced test development is still in the intuitive or artistic state. More often than not the procedure is iterative. For example, attempts to specify classes of behavior may begin by first specifying varieties of test items. These items might be subjected to behavioral analysis and behavioral class descriptions are then induced. This may lead to further specification of items or redefinition of behavior classes.

Permeating all of the discussion that follows is the notion of a theory of performance (Miller, 1962; Hively, 1970) or an analysis of the psychological processes underlying task performance. This type of process analysis is used to structure the classes of behavior defining various levels of achievement and in interpreting specific item performance as representing the class of behavior defined.

DOMAIN DEFINITION

Of the four characteristics of criterion-referenced testing[1] outlined earlier, specifying classes of behavior that define different levels of achievement is the most difficult to achieve. The failure to adequately specify this domain has led to recent criticisms of criterion-referenced testing (e.g., Ebel, 1970; Stanley & Hopkins, 1972). Since these criticisms hark back to the inadequacy of the old percentage grading system, perhaps the demise of that system was also due to the domain specification failure.

A complete exposition on domain specification is beyond the scope of this paper. It is useful, however, to sketch out some of the dimensions of the problem so that the practical developer of criterion-referenced tests may take them into consideration. These dimensions include establishing various levels of achievement, the relationship between ultimate and proximate achievement levels, the nature of the domain specification, and the derivation of domain descriptions.

Levels of Achievement

Performance or achievement criteria can be established at any convenient point in the instructional process. For example, the classes of behavior defining various levels of competence can be specified at the termination of a course, at the termination of a unit of instruction (i.e., smaller within-course segments of instruction), or at any other point during the course of instruction. The definition of these behavior domains will be guided by the nature of the instructional system and the purpose for which the information will be used, e.g., certification of attainment, within curriculum placement, or diagnosis of deficiencies (cf. Glaser & Nitko, 1971).

At the termination of instruction broad domains of performance are definable. The definition and analysis of these domains occur at several levels ranging from the definition of the desired outcomes of the entire educational enterprise, at one extreme, to the specification of the desired outcomes at the termination of a particular subject-matter course, at the other extreme. The former is likely to yield many domain definitions, be divergent, and require many tests in order to assess pupil outcomes. The latter leads to fewer domains, is more convergent in terms of outcome categories, and may result in fewer tests.

---

[1] While it may be useful for some to avoid the term criterion-referenced testing and focus on criterion-referenced score interpretation (e.g., Simon, 1969; Davis, 1970), it seems more useful to refer to "tests" in the context of this paper. In order to have criterion-referenced score interpretation, scores need to be referenced back to the behavior domain. Hence, focus in development should be primarily on the domain of behavior and the derivation of test tasks to elicit that behavior, rather than short-cutting these and focusing mainly on the scores (cf., Jackson, 1971).

Ultimate and Proximate Behavior

Defining levels of achievement at various points in instruction raises the issue of what kinds of behavior are important enough to be included in a domain specification. While this is an old area and subject to considerable debate and discussion, it is not yet resolved. The importance of the distinction between proximate and ultimate objectives of instruction for educational test developers was articulated several years ago by Lindquist (1951).

Educational practice generally assumes that the knowledge and capabilities with which the learner leaves the classroom are related to the educational goals envisioned by society. This assumption implies that the long-range goals the learners are to attain in the future are known and that the behaviors with which the learners leave a particular course actually contribute to the attainment of these goals. What is closer to reality, however, is that the long-term relationship between what the student is taught and the way he is eventually required to behave in society is not very clear (Glaser & Nitko, 1971).

In contrast to ultimate goals, proximate goals define the domains of performance that a learner displays at the end of a particular instructional situation (e.g., course or grade level). It should be noted that proximate objectives are not defined as the materials of instruction nor as the particular sets of test items that have been used in the instructional situation. For example, at the end of a course in spelling one might reasonably expect a student to be able to spell certain classes of words from dictation. During the course, certain of these words might have been used as examples or as practice exercises. The instructor would be interested in the student's performance with respect to the class or domain of words as a proximate objective of instruction and not the particular words used in instruction. Thus, to assess a student's performance with respect to a domain, *one may need to consider the transfer relationship between the items in the domain and the preceding instruction.*

General Nature of Domain Specification

The specification of the domain of instructionally relevant achievement behaviors can profit much from the suggestions for "universe specification" advocated by Cronbach (1971). As Cronbach has pointed out, too often attention is paid only to the selection of subject-matter topics. The nature of the stimulus and the description of the response are ignored. Proper domain specification requires that both stimulus and response descriptions be included. Thus,

> A proper response specification deals with the result a person is asked to produce, not the process(es) by which he succeeds or fails. 'Reads printed words aloud' is a description of an observable response; it says nothing about whether the reader is to look and say or to sound the word out. A person who insisted on separating these two response processes would have to devise a new task specification, perhaps requiring the reading of nonsense constructions that no subject has seen before. If a category of the form say, 'ability to evaluate arguments' is to mean anything as a task specification, the designation must be fleshed out to describe something about the stimulus, the accompanying injunction to the subject, and the aspect of the behavior to which the scorer is directed to attend (Cronbach, 1971, p. 454).

In this sense, use of the *Taxonomy of Educational Objectives* (Bloom, 1956) is insufficient for domain specification since the categories described therein are inferred psychological processes. However, to adequately specify the dimensions of the performances to be included in the domain, one may need to invoke a theory of performance (Hively, 1970; Miller, 1962) to decide which stimulus and response characteristics are relevant for domain description. This point will arise again when deriving tasks from the behavior description is discussed.

Derivation of Domain Description

While in practice the generation of performance domains is often ultimately tied to the actual specification of the tasks (stimuli) themselves, this derivational process is discussed separately here. It should be noted, however, that the state of the technology for determining the content and attributes of *what* is learned is not well developed, particularly where behavioral characteristics of complex school-like performances is concerned (Glaser & Resnick, 1972).

One practical method for deriving domain descriptions for smaller classes of behavior, such as a domain of behavior relevant for a unit[2] of instruction, is the procedure stemming out of Gagné's work on learning hierarchies (e.g., Gagné & Paradise, 1961). [A modification of this procedure, which seems to give more replicable results, has been provided by Resnick (Resnick, Wang, & Kaplan, 1970).] The analysis of learning hierarchies begins with any desired instructional objective, behaviorally stated, and asks in effect: To perform this behavior what prerequisite or component behaviors must the learner be able to perform? For each behavior so identified, the same question is asked, thus generating an ordered hierarchy of behaviors based on testable prerequisites. The analysis can begin at any level and always specifies what comes earlier in the curriculum. It should be noted that as it is used here, hierarchy analysis is a tool for domain definition. Whether all students' learning should progress through the hierarchy in the same way is an empirical question for instructional psychology.

---

[2] The analysis of learning hierarchies need not be restricted to units of instruction, of course. It may be possible to apply the procedure to broad curricular areas.

As a result of this type of analysis and domain specification, the test developer is provided with the essential information about what behaviors are to be observed and tested in order to determine the status of the learner with respect to the achievement continuum. Thus a hierarchical analysis provides a good map on which the attainment, in performance terms, of an individual student may be located. The uses of such hierarchies in designing a testing program for a particular instructional system are described elsewhere (Glaser & Nitko, 1971).

A serious question that can be raised is how much of education can be analyzed into hierarchical structures. The answer to the question is very much an open, experimental matter. Three things should be noted, however (Glaser & Nitko, 1971). First, the development of hierarchies for complex behaviors may lead to several such structures, each of which is "valid" with different kinds of learners, but none of which, taken alone, is valid for all learners. Second, the analysis of behaviors into components and prerequisites leads to structures that stand as hypotheses open to empirical verification. Third, in actual instructional practice there is always a functional sequence wherein the instructor has at least an intuitive hierarchy through which he proceeds.

Another point to remember is that criterion-referenced interpretations are most useful when the behavior domain has an orderly progression (Cronbach, 1970). Hierarchy analysis, or a similar procedure, would seem to be a useful tool in discovering these progressions.

The use to which the test is to be put will to a large extent determine the nature of the performance to be included in the domain definition. For example, one may develop performance domains by analysis of an "expert's" behavior or by the analysis of an "amateur's" behavior (Hively, 1970). It may well be that certain elements of performance will drop out as task proficiency increases. For assessment of initial stages of learning, therefore, it may be that more components need to be included in the domain definition (and consequently on the test) than at later stages of learning. This would seem to imply a distinction between diagnosis, placement, and final (terminal) learning assessment (see Glaser & Nitko, 1971).

DEFINING CLASSES OF ITEMS

Closely associated with the definition of behavior classes related to levels of achievement is the translation of these behavioral statements into sets of test situations—test tasks or test items. Although discussed here separately, in practice these two steps are often iterative. Performance domains tend to be verbal statements and descriptions (e.g., behavioral objectives) whereas test situation descriptions tend to be more concrete in that the characteristics of the testing situation and the various type of admissible test items are mapped out and specified. Test items here refer to any carefully described ". . . stimulus conditions under which a student is expected to respond, together with the specifications for recording and scoring his response when it occurs (Hively, 1970)." Items include both performance and traditional paper-and-pencil types of items as long as these are derived from the domain definitions.

Item Forms

A useful tool for criterion-referenced tests is item forms analysis (Hively, 1966; Hively, Maxwell, Rabehl, Sension, & Lundin, 1973; Hively, Patterson, & Page, 1968; Osburn, 1968). Item forms analysis is a variation on task analysis. It is the process whereby behavioral statements are analyzed in order to derive classes of items which elicit the various aspects of the behavior class. As a result of this analysis, one or more item forms are derived for each behavior class. An item form consists of a specification of the invariant part of the class of items together with (a) an indication of which parts of the items are variable, (b) a specification of elements which can be used in the variable parts of the items, and (c) a specification of the rules by which one selects an element from the set of variable elements to derive a particular item (Hively, 1970; Hively, et al., 1973). The variant part of the item is called a *shell;* the sets of elements which can be used in the variable parts are called *replacement sets;* and the rules by which one samples from the replacement sets are called the *replacement structure* (Hively, 1970; Hively, et al., 1973).

In practice, one often cannot go directly from a verbal statement of a behavior class to an item form. The procedure usually is to first develop prototype items admissible as test tasks under the described behavior. Process and component analysis (cf. Resnick, Wang, & Kaplan, 1970) of these prototype items often leads to a modification of the original behavior specification, elimination of some of the prototype items as not implied by the behavior class, or a rewriting of the prototype items. In examining these prototype items to determine their fit to the behavioral definition one invokes a behavioral analysis and a theory of performance. This process involves more than superficial judgment and sorting. The questions that need to be answered are: (1) Does this item contain the stimulus characteristics implied by the behavioral statement? (2) Will the examinee's response to this item be indicative that he indeed has the desired response in his repertoire?

Once the set of prototype items has been delineated item forms can be induced. The prototype item is one member of the class of items implied by an item form. The task here is to identify the general form (format) of the items, the item shell, the variable elements, and the admissible replacement sets. Again, this implies a behavioral analysis and a theory of performance.

Item Tryout Data

As part of the procedure for defining test tasks that are consistent with domain definitions, it is necessary to establish empirical procedures for tryout of items. A major purpose of traditional item-tryout procedures is to collect data necessary to improve the test items. This is no less true when criterion-referenced test items are developed.

Tryout of items for criterion-referenced test development seeks to further refine and polish the domain of test tasks. All the ambiguities that are inherent in traditional item writing are inherent in criterion-referenced item writing. Further, since item forms are developed using behavioral analysis and performance theory, the data from item tryout are used to check on the adequacy of this initial analysis. Often this will lead to a respecification of the item form or one or more of its components —replacement sets or replacement structure (cf. Osburn, 1968).

There are those who advocate either explicitly (e.g., Stenner & Webster, 1971) or implicitly (e.g., Baker, 1971) that items designed to test a specific class of behaviors be homogeneous. Homogeneous tends to be defined in terms of item and total test score parameters such as discrimination indices and internal consistency reliability estimates. These correlation-related indices tend to be maximized when each item measures the same factor (process) (Lord, 1958). The insistence on homogeneity in this sense is too sweeping and is poor psychology. It leads to statistical techniques being used to drive the definition of performance domains. There is no logical basis for contending a priori that any domain of performance identified as instructionally relevant ought to be homogeneous (cf. Cronbach, 1971). Homogeneity should be viewed as a question for empirical experimentation and item performance theory (cf. Bormuth, 1970) and would probably vary with the target population and the class of behaviors under consideration. Heterogeneity would mean that a larger number of observations are needed before adequate generalizations about domain performance can be made.

Hierarchy Validation

If hierarchy analysis is used to develop the test domain, empirical data needs to be collected to validate this structure in terms of the items defining the various levels of the hierarchy. One should distinguish what might be called the "psychometric" hierarchy[3] from the learning hierarchy.

---

[3] For an example of procedures used to validate psychometric hierarchies see Wang, Resnick and Boozer (1971) and Ferguson (1970).

Classes of test tasks (items) can be ordered in hierarchical ways which may bear little relationship to the sequence in which learning should proceed. If the hierarchical ordering of the domain implies an instructional sequence, or if it represents a hypothesis about behavioral acquisition derived from instructional theory, then empirical transfer studies are required as well. Thus, criterion-referenced testing is not exempt from construct validation studies (cf. Cronbach, 1970).

Item Performance and Instruction

An important consideration in the tryout of test items in this context is the relationship between instruction and the test item domain. The tryout data is dependent on: "(1) the characteristics of the item itself, (2) the program of instruction with which it is associated, (3) the sample of the students from whom the data were collected, and (4) the conditions under which the students worked (Hively, 1966, p.7)." These are factors which influence the interpretation of tryout data and the subsequent decisions that are made concerning item and domain revision.

If the behavioral domain and subsequently derived item classes are based on some inferred process (e.g., application in the Bloom *Taxonomy*) or an inferred psychological construct (e.g., a hierarchy of prerequisite behaviors), then the content and nature of the examinees' previous learning history (i.e., instruction) need to be considered in interpreting tryout data. A similar point is made by Bormuth (1970) who calls for the development of procedures for relating the structure of the items to the structure of the instruction. For example, to adequately derive classes of test tasks measuring transfer, application, and evaluation behaviors it is necessary to eliminate from the item form those items on which the students were given practice, thus leaving those items that elicit responses not explicitly taught, but which can be deduced from instruction. Without such procedures, it is not possible to determine whether the classes of items are indeed achievement items, as opposed to general knowledge or aptitude items.

The development of items for criterion-referenced tests and the associated empirical data generated by tryout and study of these classes of items seem to call for aspects of achievement test theory that are as yet not well developed. Bormuth labels these *item-writing theory* and *item-response theory*. Item-writing theory would lead to the development of procedures for defining classes of items (item forms) and item-response theory would lead to explanations of the processes that account for responses to classes of items. The developer of criterion-referenced tests should refer to Bormuth's book for suggestions along these lines and for indications of some of the problems involved in pursuing research in these areas. It should be emphasized that theories and research in these areas are currently inadequate or completely lacking.

SELECTING ITEMS TO APPEAR ON THE TEST

Once the behavior domain and the classes of items have been specified the final stages of test development can proceed. It might be argued that the preceding discussion concerning domain definition is no more than what any test developer should do in order to maximize content validity, regardless of whether a criterion-referenced or a norm-referenced test is to be developed. While this is probably more of a fond hope than a reality, one is still inclined to agree that perhaps all test developers should take such care in developing tests. It should be noted, however, that content validity implies an indication of the sampling plan by which the particular items that appear on a particular test form are selected from the domain of all items (Cronbach, 1970).

It is assumed here that empirical data and performance theory support the definitions of achievement levels in the domain and the classes of test tasks operationalizing these behavior classes. The task is to select items to put on a form of the test in such a way that performance on that test will be a basis for an inference about the examinee's performance in the domain. It has already been mentioned that criterion-referenced test score interpretation is most meaningful when the behavior domain has an orderly progression. This implies taking advantage of the psychological structure of the subject-matter domain in selecting test items.

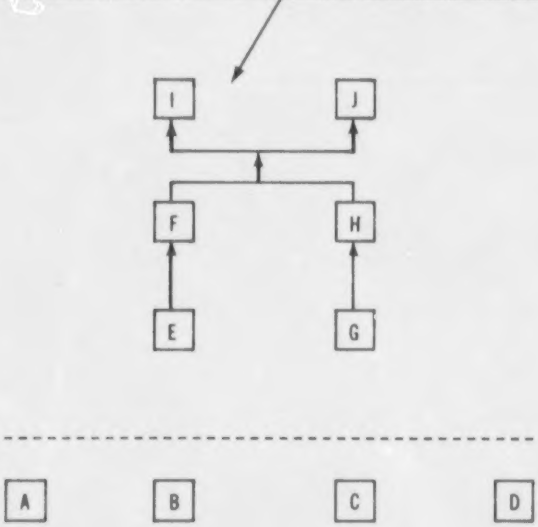
Examples of Item Selection for Curriculum Placement

If an instructional system is adaptive, it will avoid teaching the student that which he has already learned and will instead offer him new goals to learn. Information is needed to answer the question, "Where in the instructional sequence should the student begin his study?" Tests built to provide this information are specific to the content and psychological structure of the particular course of instruction with which the student is faced.

In broad areas such as an entire course or an entire curriculum area, neat hierarchies of the Gagné type covering the entire course of instruction may not exist or may become very complicated. Nevertheless, some sequencing of instructional objectives is possible. An illustration of this is shown in Figure 1 in which an elementary school mathematics curriculum has been defined in terms of approximately 350 instructional objectives. The content has been broken down into ten topics which are roughly in a prerequisite order (from top to bottom in the figure). Further, each topic has been developed over a range of complex behaviors that are also in a rough prerequisite order (from Level A through Level G in the figure). Each cell of the grid represents several instructional objectives and is called a unit of instruction. The objectives in a unit of instruction can usually be arranged in a hierarchy that leads to a few terminal goals for that unit. The inset shows (hypothetically) how a short sequence of objectives might look for one unit of instruction. Within a single unit, in general, there will be prerequisite behaviors from earlier topics and lower levels. These are labeled as behaviors A, B, C and D in the inset.

One way to place a pupil in this curriculum is to develop a two-stage

Content (Topic)	Level of Complexity						
	A	B	C	D	E	F	G
Numeration/Place Value	*	*	*	*	*	*	*
Addition/Subtraction	*	*	*	*	*	*	*
Multiplication		*	*	*	*	*	*
Division		*	*	*	*	*	*
Fractions	*	*	*	*	*	*	*
Money	*	*	*	*			
Time	*	*	*	*	*		
Systems of Measurement		*	*	*	*	*	*
Geometry		*	*	*	*	*	*
Applications		*	*	*	*	*	*



*Indicates a unit of instruction consisting of one or more instructional objectives.

Figure 1. Example of Curriculum Layout for Individually Prescribed Instruction Elementary Mathematics

MATHEMATICS PLACEMENT PROFILE

Name John Smith **Date** 5/70 **Grade** 5

School Sweetdate **Teacher** Mrs. Jones **Room** 12

Mathematics Area	Placement Level A-G							Placed at Level
	A	B	C	D	E	F	G	
Numeration/Place Value	▨	▨	▨	▨				E
Addition/Subtraction	▨	▨	▨	▨	▨			F
Multiplication	▨	▨	▨	▨				E
Division	▨	▨	▨					D
Fractions	▨	▨	▨					D
Money	▨	▨	▨	▨				--
Time	▨	▨	▨	▨	▨			--
Systems of Measurement	▨	▨	▨	▨	▨			F
Geometry	▨	▨	▨	▨				E
Applications	▨	▨	▨					D

Figure 2. Example of Placement Profile for a Hypothetical Student with Respect to the Mathematics Curriculum of Individually Prescribed Instruction

placement test (Cox & Boston, 1967). The first-stage test is broad-ranged over the curriculum. The results are used to place a student at a unit in each topic or content area. The second-stage test is narrow-ranged and tests the domain of behavior implied by a single unit. The results are used to place a student at a particular objective within a unit. The first-stage test needs to be administered only once at the beginning of a course of study. After completing instruction on the first unit of study, the student is given the second-stage test for the next sequential unit. Thus, he is placed at each successive unit in the curriculum. Figure 2 shows a completed first-stage placement profile for a hypothetical student. Figure 3 shows what a completed second-stage placement profile might look like.

The broad-range test is actually a battery of tests consisting of one test for each topic. Each subject would predict for each topic the last unit in the sequence from A to G in which the student would be successful. Traditional item-selection procedures that seek to maximize predictive validity would seem appropriate for this type of broad-range test. If the behaviors defined within a unit are hierarchical, then one could select

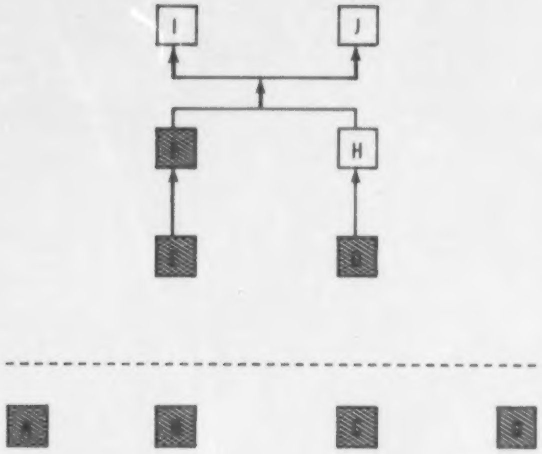


Figure 3. Placement Profile for a Hypothetical Student (Shaded boxes mean that the student has sufficient mastery of these instructional goals to proceed with a new instructional goal.)

items from the domains that define the terminal objectives for that unit, and depend on the prerequisite nature of the hierarchy to subsume the other behaviors in the unit. If a within-unit hierarchy does not exist, then selecting items from the domains of all the within-unit behaviors would seem to be required. Care should be taken, however, in using correlational indices for this type of prediction; it is the absolute level of attainment of unit skills that is of prime importance.

The second-stage type of unit test serves as another example of how items might be selected by taking advantage of the psychological structure of the subject-matter content. If the unit behaviors are hierarchical and domains of items are defined for each node in the hierarchy, then a branched test can be used to obtain a pupil's profile with respect to this hierarchy. Thus, if an examinee was successful on items testing one objective in the hierarchy, this would indicate that items from earlier objectives in the hierarchy would be passed as well.[4] Procedures for branched testing initially proposed by Ferguson (1970) and further elaborated by Hsu (Ferguson & Hsu, 1971; Hsu & Carlson, 1972) have been successfully used in an elementary mathematics curriculum when coupled with item forms and a computer.

---

[4] Such elaborate procedures would have to be balanced out against efficiency criteria. For example, in small hierarchies consisting of a few nodes a tailored test would be more elaborate than necessary. A student might be placed more quickly and efficiently by simply testing all nodes.

Figure 4 is a schematic illustration of terminal and prerequisite instructional objectives for an addition-subtraction unit from the elementary arithmetic curriculum of the Individually Prescribed Instruction Project (Lindvall & Bolvin, 1967). Each box represents one objective. The objectives are arranged in a branched hierarchy. Objectives 6, 17, and 18 are terminal objectives for the unit; the remaining objectives are prerequisites. Each of these prerequisites and terminal objectives is defined by one or

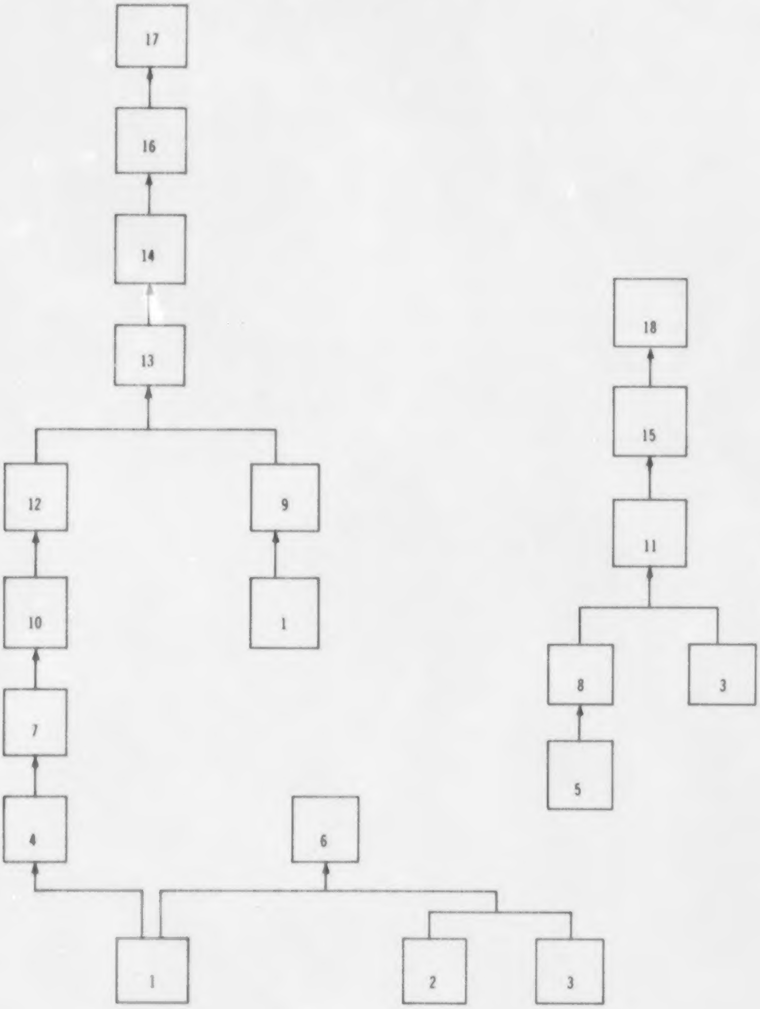
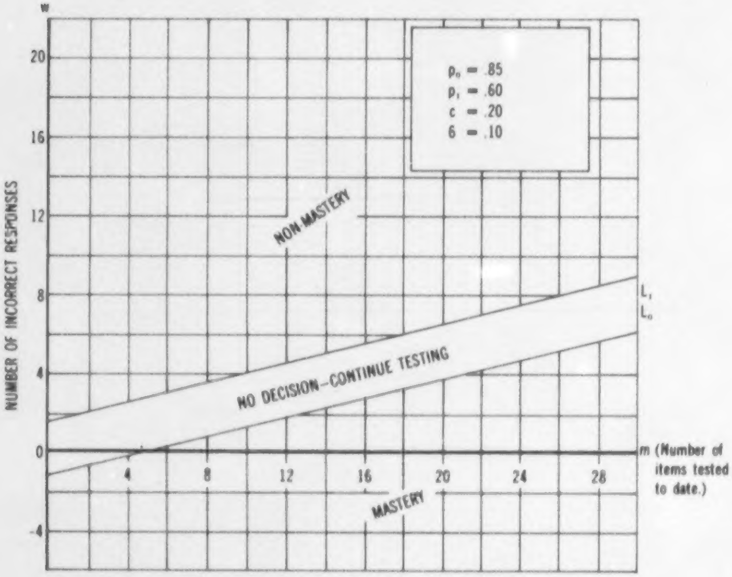


Figure 4. An Example of a Hierarchy of Skills in an IPI Mathematics Unit

more item forms which are then programmed for use on the computer. The testing is done on an individual basis at a computer terminal.

The object of the testing scheme is to locate a pupil at one of these objectives or "boxes" as quickly as possible and in such a way that he demonstrates mastery of objectives below his location and non-mastery of objectives above his location. The decisions for which the testing procedure must provide information are (1) what objectives should be tested and (2) whether the pupil has mastery or non-mastery[5] of the objectives that are tested. A decision needs to be made about every objective, but the trick is to make these decisions without testing every objective, and to minimize the testing for those objectives that are tested.

On this basis, a set of decision rules is devised that combines the capabil-



H_0: p = .85 (Student has sufficient mastery, omit instruction)
H_1: p = .60 (Student does not have sufficient mastery, give instruction)

Figure 5. Graph Illustrating Sequential Probability Ratio Test for Determining Whether a Student Does or Does Not Need Instruction on an Objective (Modified from Ferguson, 1970)

---

[5] By mastery it is meant that ". . . an examinee makes a sufficient number of correct responses on the sample of test items presented to him in order to support the generalization (from this sample to the domain or universe of items implied by an instructional objective) that he has attained the desired, pre-specified degree of proficiency with respect to the domain (Glaser & Nitko, 1970, p.641)."

ities of the computer with statistical logic and subject-matter logic. This allows "on-line" decisions to be made about what is to be tested and how extensively it is to be tested. The procedure breaks away from the traditional "test now, decide later" schemes that have received recent criticism (e.g., Green, 1969).

A decision about mastery of one objective can be made by using the sequential probability ratio (Wald, 1947). An example of the situation is shown in Figure 5. The test length varies from pupil to pupil. A pupil is given only as many randomly-selected test items as are necessary to make a mastery or non-mastery decision with respect to a fixed mastery criterion and with prespecified Type I and Type II error rates. After each item is administered and scored, a decision is made to declare mastery, continue testing, or to declare non-mastery. With the number of items a random variable, it is possible, in this example, to make a mastery decision with as few as 6 items and a non-mastery decision with as few as 2 items. Not all mastery and non-mastery decisions are made this quickly; it depends on the response pattern of the pupil.

Figure 5 illustrates the procedure for one objective. The problem that remains is that a decision needs to be made about every objective. Since the objectives are organized into a prerequisite sequence, the sequence itself can be used in the decision-making process. This results in the compound *branching rule* shown in Table 1 for determining the next objective to be tested. The "next objective to be tested" depends on whether the student is declared a master or a non-master *and* on his response pattern that led to this decision. This is illustrated by the arrows sketched on Figure 6.

Table 1. Branching Rules for Computer-Assisted Placement Testing

Decision for 1 Skill	Pupil's Response Data (p)	Branching Rules (**Next** Skill to be Tested)
Mastery ($p \geq .85$)	HIGH ($p \leq .93$)	Branch **up** to **highest** untested skill.
	LOW ($.85 \leq p \leq .93$)	Branch **up** to skill **midway** between this skill and highest untested skill.
Non-Mastery ($p \leq .60$)	HIGH ($.43 \leq p \leq .60$)	Branch **down** to skill **midway** between this skill and lowest untested skill.
	LOW ($p \leq .43$)	Branch **down** to **lowest** untested skill.

Testing begins at an objective in the middle of the hierarchy and continues until the branching rule cannot be satisfied. At that point, the objective tested is the proper location of the student in the hierarchy.

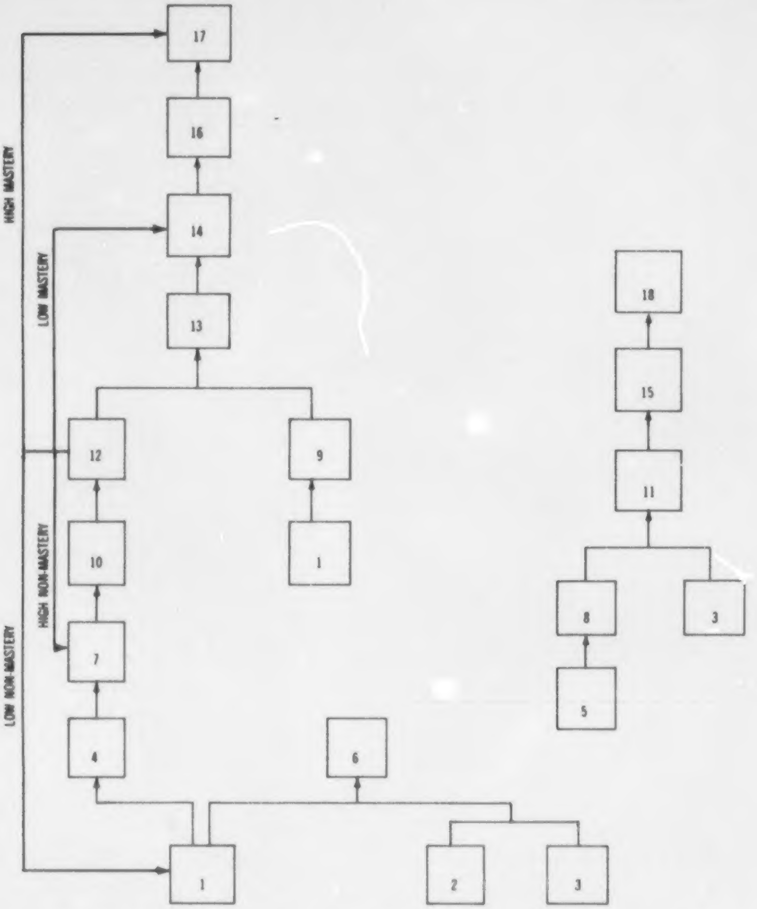

HIGH MASTERY
LOW MASTERY
HIGH NON-MASTERY
LOW NON-MASTERY
17
16
14
13
12
9
10
1
7
4
6
1
2
3
18
15
11
8
3
5

Figure 6. An Example of the Application of the Branching Rules of Table 1 to the IPI Mathematics Unit in One Instance

(Note: Only one of the "arrows" would be followed to locate the next objective to be tested. The branching rules would be reapplied after testing the next objective.)

Untested skills can be assumed mastered or unmastered according to their position in the hierarchy and the student's response data.

An individual's testing session results in a profile similar to the one shown earlier in Figure 3. The student would begin his instruction in this unit on the next sequential objective that was unmastered.

Elaborations on how items are selected and generated from item forms by the computer are given elsewhere (Ferguson & Hsu, 1971; Hsu &

Carlson, 1972). Figure 7 is a flow chart that illustrates the item selection, administration, scoring, and decision-making procedures in the testing situation. It should be noted that this type of criterion-referenced branched testing is still in the developmental stage and that evidence concerning its appropriateness needs to be provided before it can be strongly recommended.

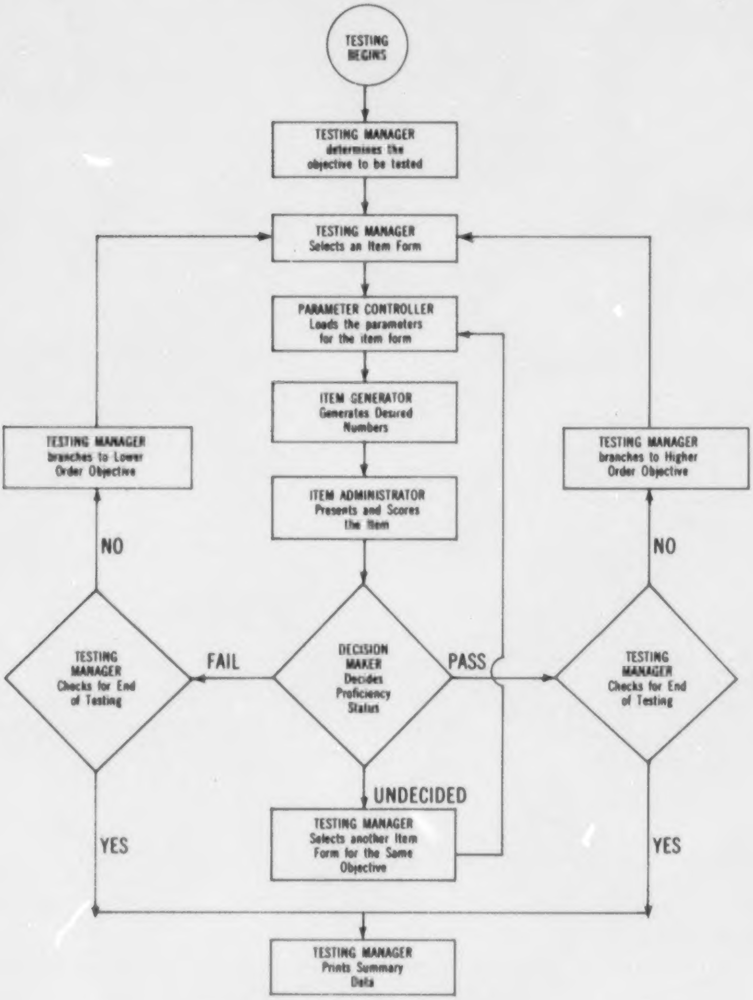

TESTING BEGINS
TESTING MANAGER determines the objective to be tested
TESTING MANAGER Selects an Item Form
PARAMETER CONTROLLER Loads the parameters for the item form
ITEM GENERATOR Generates Desired Numbers
ITEM ADMINISTRATOR Presents and Scores the Item
TESTING MANAGER branches to Lower Order Objective
TESTING MANAGER branches to Higher Order Objective
NO
NO
TESTING MANAGER Checks for End of Testing
FAIL
DECISION MAKER Decides Proficiency Status
PASS
TESTING MANAGER Checks for End of Testing
UNDECIDED
TESTING MANAGER Selects another Item Form for the Same Objective
YES
YES
TESTING MANAGER Prints Summary Data

Figure 7. Execution Model for Pretests and Posttests Using Item-Cluster Generators (Adapted from Ferguson & Hsu, 1971)

CRITERION-REFERENCED TEST SCORES

Criterion-referenced test scores lead to an inference about the performance characteristics of the examinee. Such scores indicate the behaviors the examinee can exhibit with respect to a defined domain of behaviors. These scores are derived scores in the sense that their interpretation is based on the psychological structure underlying the behavior domain.

In the examples illustrated in figures 2 and 3, the unit of instruction and the node in the hierarchy are defined by classes of behaviors. A particular score on the geometry subtest, for example, might mean that the examinee can perform all lower-level behaviors up to and including: identifying pictures of open continuous curves, lines, line segments, and rays; stating how these are related to each other; writing symbolic names for specific illustrations of them; identifying pictures of intersecting and non-intersecting lines; and naming points of intersection. The score would also mean that the examinee could *not* demonstrate higher-level behaviors.

Scores may also be related to expectancy tables, thus indicating the probabilities associated with various score-behavior class performance combinations (Cronbach, 1970). This would combine norm-group data with performance data and aid in the overall interpretation of performance not tested. For example, relating acquired levels of performance to chances of being successful in new instructional situations broadens the interpretation of criterion-referenced scores. Obviously, normed-referenced scores such as percentile ranks, standard scores, grade equivalents, and so on can be obtained from criterion-referenced tests as well.

An issue often closely associated with criterion-referenced testing is that of mastery learning and mastery testing. A full discussion of mastery testing is beyond the scope of this paper. The reader is referred to papers by Bloom, Hastings, and Madaus (1971), Block (1972), Bormuth (1971), Ebel (1970), and Glaser and Nitko (1971), for some discussion of this problem as it relates to testing. It is noted here that a criterion-referenced test does not necessarily imply flawless performance nor that any examinee necessarily meet a given standard of competence. What is implied, however, is the notion that such levels of competency be defined in terms of performance (Nitko, 1970).

INSTRUCTIONAL SYSTEMS AND TESTING

It is important to point out that the kinds of tests that are developed and used will depend on the decision framework within which the test-provided information is employed (e.g., Cronbach & Glaser, 1965). It has been indicated that criterion-referenced tests will probably find their greatest use in instructional situations. Since there are a variety of ways in which instructional systems can be designed and operated to adapt to individual differences (Cronbach, 1967), the design of testing programs needs to take the instructional system into account. This means that various mixtures of criterion-referenced and norm-referenced test varieties will be needed depending on the particular instructional system. Thus, in the overall planning and designing of a testing program, decisions about when (and whether) criterion-referenced tests are to be used need to be made.

One example of how criterion-referenced and other types of test information can be designed into a particular kind of individualized instructional system has been given by Glaser and Nitko (1971). The discussion there indicates how the various kinds of instructional decisions that need to be made are determined as well as the kinds of tests that need to be developed to provide this kind of information. Similar analyses of other types of instructional systems need to be made and testing programs need to be developed in the context of these analyses.

SUMMARY

This paper has reviewed the requirements for the construction of criterion-referenced tests that would be used in instructional situations. It has tried to indicate the problems faced in the practical construction of such tests and some of the techniques that have been found to be of some value in solving these problems. Adequate solutions do not exist for all of the problems raised. In particular, procedures are needed for the solution of the following problems:

1. Defining the behaviors to be taught and tested for in the instructional situation.
2. Task analysis as it relates to school-like behaviors.
3. Relationship between what is tested and the ultimate objectives of the individual and society.
4. The relationship between the behavioral domain and the domain of tasks serving as the potential item domain.
5. Specification of the domain of tasks in terms of their stimulus and response characteristics.
6. The ordering of the domain of behaviors in terms of their psychological structure.
7. Data related to the generalizability of samples of behavior to the behavioral domain.
8. Construct validation of proposed orderings of the behavioral domain.
9. The development of an item-writing theory and an item-response theory.
10. Development of procedures for determining mastery of identified behavior.

While solutions to the above problems would lead to improved criterion-referenced test construction practices, it should not be assumed that criterion-referenced information is all that is needed to make instructional decisions. Without an analysis of the kinds of instructional decisions that need to be made in a given instructional situation, discussions about tests, testing procedures, and test development tend to be fruitless.

REFERENCES

Baker, E.L. The effects of manipulated item writing constraints on the homogeneity of test items. *Journal of Educational Measurement*, 1971, *8*, 305–309.

Block, J.H. Student evaluation: Toward the setting of rational, criterion-referenced performance standards. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Bloom, B.S., et al. (Eds.) *Taxonomy of educational objectives, handbook I; Cognitive domain*. New York: David McKay, 1956.

Bloom, B.S., Hastings, T.M., & Madaus, G.F. *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.

Bormuth, J.R. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

Bormuth, J.R. Development of standards of readability: Toward a rational criterion of passage performance. Final Report, USDHEW, Project No. 9-0237. Chicago: The University of Chicago, 1971.

Bruner, J.S. *Toward a theory of instruction*. Cambridge, Mass.: The Belknap Press of Harvard University Press, 1966.

Cox, R.C., & Boston, M.E. Diagnosis of pupil achievement in the Individually Prescribed Instruction Project. Working Paper 15. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, 1967.

Cronbach, L.J. How can instruction be adapted to individual differences? In R.M. Gagné (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles E. Merrill, 1967.

Cronbach, L.J. *Essentials of psychological testing*. (3rd ed.) New York: Harper and Row, 1970.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Cronbach, L.J., & Glaser, G.C. *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1967.

Davis, F.B. Criterion-referenced tests. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Ebel, R.L. *Content-standard test scores. Educational and Psychological Measurement*, 1962, *22*, 15-25.

Ebel, R.L. Some limitations of criterion-referenced measurement. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Ferguson, R.L. A model for computer-assisted criterion-referenced measurement. *Education*, 1970, *81*, 25–31.

Ferguson, R., & Hsu, T.C. The application of item generators for individualizing mathematics testing and instruction. Publication 1971/14. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, 1971.

Flanagan, J.C. Units, scores, and norms. In E.F. Lindquist (Ed.), *Educational measurement*. Washington: American Council on Education, 1951.

Gagné, R.M., & Paradise, N.E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs*, 1961, *75*.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, *18*, 519–521.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971, 625–670.

Glaser, R., & Resnick, L.B. Instructional psychology. *Annual Review of Psychology*, 1972, *23*, 207–276.

Green, B.F. Comments on tailored testing. In W. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. New York: Harper and Row, 1969.

Harris, M.L., & Stewart, D.M. Application of classical strategies to criterion-referenced test construction: An example. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hively, W. Preparation of a programmed course in algebra for secondary school teachers: A report to the National Science Foundation. Minnesota State Department of Education, Minnesota National Laboratory, 1966.

Hively, W. Domain-referenced achievement testing. A paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project*. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hively, W., Patterson, H.L., & Page, S. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1963, *5*, 275–290.

Hsu, T.C., & Carlson, M. Oakleaf school project: Computer-assisted achievement testing. Technical Report. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, February, 1972.

Jackson, R. Developing criterion-referenced tests. ERIC/TM Report 1. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1971.

Lindquist, E.F. (Ed.) *Educational measurement*. Washington: American Council on Education, 1951.

Lindvall, C.M., & Bolvin, J.O. Programmed instruction in the schools: An application of programming principles in "Individually Prescribed Instruction." In P. Lange (Ed.), *Programmed Instruction*, 66th Yearbook, *Part II*. Chicago: National Society for the Study of Education, 1967, 217–254.

Livingston, S.A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13–26.

Lord, F.M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, 1958, *23*, 291–296.

Miller, R.B. Task description and analysis. In R.M. Gagné (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart, and Winston, 1962.

Nitko, A.J. Criterion-referenced testing in the context of instruction. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Osburn, H.G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, *28*, 95–104.

Resnick, L.G., Wang, M.C., & Kaplan, J. Behavioral analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. Monograph 2. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, December, 1970.

Simon, G.B. Comments on "Implications of criterion-referenced measurement." *Journal of Educational Measurement*, 1969, *6*, 259–260.

Stanley, J.C., & Hopkins, K.D. *Educational and psychological measurement and evaluation*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Stenner, A.J., & Webster, W.J. Educational program audit handbook. Arlington, Va.: The Institute for the Development of Educational Auditing, 1971.

Wald, A. *Sequential analysis.* New York: Wiley, 1947.

Wang, M.C., Resnick, L.B., & Boozer, R.F. The sequence of development of some early mathematics behaviors. Publication 1971/6. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, 1971.

PROBLEMS OF OBJECTIVES-BASED MEASUREMENT

Chester W. Harris
University of California, Santa Barbara

Our immediate task is to take the materials in the preceding papers, all produced by people working on the development and application of objectives-based measurement, as a starting point for an organization of problems with the intent of explicating the nature of these problems and their importance. If this task is well done it should assist us in going beyond a simple identification of problems of objectives-based measurement and moving toward a specification of what can be done to solve or possibly transform these problems. Problem solving may involve two kinds of attack: one, finding existing solutions that can be adapted or applied; and two, creating new solutions. The identification of existing solutions may turn out to be richest in usable result, but we should also hope for new solutions, possibly created in the near future and by some of us, and possibly created later and by others who follow and are made aware of our efforts here.

In attempting to form a structure as a way of organizing the problems and concerns that we have identified, I became aware of certain characteristics of the task. First, the dimensions of the attributes or characteristics of the problems probably do not form a set of completely crossed factors, and an attempt to force such neatness did not seem to be productive. Instead, it seemed wise to recognize that vacant cells may exist and that nesting may be a more realistic view than crossing. Second, the problems are characterized—at least to some extent—by a sequential mode and thus have something like a tree structure when viewed from one standpoint. Attention to ordering and partial ordering as structural principles seemed to be required. Third, the problems themselves seemed to touch existing bodies or "scraps" of theory at a number of points, and I felt that reformulating problems in terms of underlying distinctions and relationships that have an abstract structure which has proved potent in the past is likely to be a profitable line of analysis. However, in following such a line of analysis it is very important to keep the direction of influence in hand, so that the problem suggests the principle—if it exists. When it does not exist, new principles are called for.

It is relatively simple to identify four questions that give an outline of the problems associated with objectives-based measurement. These questions are:

I wish to thank James Block, Rodney Skager, and Melvin Novick for their comments on an earlier draft of this paper. It is always a pleasure to receive incisive observations that are designed to strengthen one's work. I hope I have used their comments effectively.

What objectives are to be reached?
How are these objectives to be written or formed in order to provide bases for instructional development, and/or bases for measurement procedures?
How are the measurement procedures to be developed?
How are the measurement procedures to be used?

Although these questions have a logical sequence, it is clear that attempts to answer any one of them can influence the answers to a different, possibly "earlier," question. Thus in practice an iterative procedure that moves back and forth among these questions may be a more realistic description of how objectives-based measurement procedures are developed than is the sequential outline just listed. I recognize this, but I also find the questions useful in pointing to clusters of problems. Let us look at some of these.

Let us assume that our concerns both begin and end with objectives. It is possible to take one or more objectives as given, and consequently to avoid the problem of validating the selection of objectives in terms of criteria such as importance and comprehensiveness. The resulting measurement procedures or tests then rest on the assumptions represented by the given objectives, and these assumptions will determine at least in part their characteristics and their uses. If, however, the objectives are not given a priori—or possibly inherited—then selection of objectives is necessary and a validation of these selection procedures is required. Two approaches to the validation of objectives can be identified. Both approaches require data about the importance and the comprehensiveness of the objectives, but they differ in the source of the data and the form in which the data are reported. One approach derives the data primarily from the study of society and reports the data to a considerable extent in terms of the judgments of competent persons. The other derives data from the study of learning networks or hierarchies and reports them in terms of propositions about prerequisites, necessary and sufficient stages, and the like. The former appears to be a validation procedure most appropriate for achievement inquiries that are *not* linked to specific instructional programs; the latter appears to be a validation principle most useful when questions about achievement are to be answered within the context of a particular instructional plan.

A distinction I see as an important one is whether or not the tests or measurement procedures are to be explicitly integrated into a specific instructional system. It is conceivable that for both cases one might begin with essentially the same objective or objectives and develop very much the same test. On the other hand, the test that is linked to the instructional system may adopt an "instructional bias" that is at least suggested by Bormuth's advice (1970) and seems to be implicit in some of the discussion by Baker and by Nitko. This "instructional bias" is illustrated by the distinction between testing to see whether or not the student read a particular book and testing to see what the reading of the book did to the student. It can probably be maintained that every test is designed to detect the experimental history of the student, and in that sense every test has an "instructional bias;" but it is the test that is designed *primarily* to detect whether or not the student read these particular books, did these particular exercises, etc., that I call the one with "instructional bias." A problem, then, for tests that are designed to be integrated into a specific instructional system, is that of how much "instructional bias" is to be either allowed or demanded. This question is not unrelated to the question of how to test for mastery of a terminal objective in a hierarchy for which there are several equally valid paths to that objective. The test may be "path dependent" to a greater or lesser extent, and the question of how much "path dependence" is to be allowed or demanded is an important one.

This recognition that selection of objectives, their definitions and elaborations, and the construction of measurement procedures may be done either within or apart from a particular instructional system suggests that our notions about types of tests should reflect these distinctions. There are possibly four types of tests or measurement procedures, each of which is required in the operation of a particular instructional program. These four types of tests exist in a two-by-two design: With respect to individual students it is desirable to have diagnostic-placement tests that give information about an individual that is relevant to the instructional system that he is to work through in order to change his behavior. This is one type of test and is designed to indicate where in a program a student can profitably begin, what characteristics he has that make one path more appropriate for him than another, etc. Still focusing on the individual student, a second type of test is the one that gives information about how well he did—a mastery test if you wish. This second type describes exit behaviors for a program or segment of program. The other two types of test are focused on the instructional program rather than on a particular student. The formative test—to borrow a term—is designed to give indications of the points at which the instruction is not working well and needs modification; it too is diagnostic but not in the sense of diagnosing an individual's difficulties. Finally, there is a summative test that is appropriate for determining how well the program brought about the intended outcomes. Some of the interesting technical questions about tests can be formulated as questions about relations among these four types of tests.

The validity of an objective is not independent of the form in which the objective is cast, including the level of detail associated with it. This seems to be true both for the role played by objectives in providing bases for instructional development and in providing bases for measurement procedures. An especially important consideration for measurement purposes is that of specifying the behaviors that are taken as the evidence of achievement of the objective. I should like to emphasize what I call a "can do, does do" distinction as a major classification of these behaviors. It is evident that most of the achievement testing in schools focuses on the "can do" class of behaviors: the evidence of the achievement of the objective is that the student can do this and this and this. These are the knowledge, skill, ability types of objectives. But there may be other objectives, the achievement of which is evidenced by what the student typically does do. These are the attitude, interest, cognitive style types of objectives—perhaps more affective than cognitive. An important point is that both types appear to demand similar specifications in order to provide bases for measurement procedures. For both it is necessary to specify what the student does, with what materials, under what conditions, etc., that is taken as the evidence of achievement of the objective. "What is the critical evidence?" is the question whose answer defines the appropriate test or observation schedule item or sets of items. We recognize, of course, that we may subsequently establish and then use a substitutive principle which takes one behavior as an index of another; this is illustrated by the common practice of using recognition behaviors as substitutes for production behaviors.

At this stage we have identified a validity component of a particular item (task) or specific observation that is not a function of a student's response nor of the relation of this response to other responses. Instead, the component is given by the logic of the definition of the critical behavior. This component is necessary for validity but may not be sufficient, since we may subsequently find that the item does not function as our logic had led us to expect. Note that the problem of validating learning sequences or networks is similar in character. If we let specific objectives label the nodes of the network, then the logic of the behavioral definition of each objective is critical for the validity of the network itself. However, good definitions may not be sufficient since the behaviors may not function as anticipated. In addition to a definitional component of validity for any item, there is a response component. It is the study of response components that has led to the bulk of what we think of as test theory and its many applications in the form of item analysis procedures, reliability estimation procedures, etc. Before we look at some of these aspects of objectives-based measurement, let us say a bit more about test items.

For the test developer an important question is whether or not it is reasonable to postulate a population of similar items for any specific objective. The critical word is "reasonable." If the answer is no, then the developer is forced to use a single-item test to measure that objective. (He may, of course, decide to aggregate several such single-item tests rather than score them separately, but this does not alter the fact that he cannot sample from a defined population of items to measure *that* specific objective.) The unique- or single-item test poses many familiar problems. Much experience indicates that it may not be very dependable as a diagnostic or placement device, as an indicator of mastery, or as a predictor of success in subsequent instruction. In addition, the single-item test, when used as an indicator of mastery, has the further limitation of providing information only about the student's ability to perform this very specific task, thus offering no evidence about his ability to perform a class of behaviors or perform the task over a range of contents.

In contrast, when it is reasonable to postulate a population of similar items for any specific objective, a sampling (of items) procedure can generate randomly equivalent tests, each of which provides an unbiased estimate of the proportion of the population of items that a given student can perform or answer correctly. The item form notion of Hively and his associates (1968, 1973) illustrates methods of defining such a population of items and of writing the rules for constructing such items. For example, the addition of two two-digit numbers can be designated as a population of items, with the nine digits plus zero being used as the replacement sets to be sampled to yield a large number of different items. Such a population is defined in formal terms, without reference to student data. Another example is the population of content-standard vocabulary items described by Ebel (1962); the rules for writing the item are specified and then the replacement set of all the words in a specified dictionary can be sampled randomly to create a test of a specified length. Again, the population is defined in formal terms and is not stratified or characterized with respect to student response data.

Past practice seems to consist primarily of defining a population of test items in terms of a fixed task and variable content. This is a correct description of the Ebel vocabulary items, in which the task is fixed (matching an English word to a definition) and the content varies over a defined population of words. It is also a possibly correct description of the population of items testing addition of two two-digit numbers: here the task is fixed and the content (the digits) varies. The logical possibility of defining a population of items by fixing the content and varying the tasks is more difficult to illustrate; apparently task tends to have a prior claim in our definitions of item populations. A third possibility is to conceptualize an item population as a completely crossed design in which each of several tasks is matched with each of several contents; this is illustrated in a recent study of concept attainment abilities (Harris & Harris, 1973). For each of these three types of populations of items, a set of items can be selected randomly to make up a test, the score on which estimates the proportion or percentage of the population of items that is known by a given student. Such a test has the obvious characteristic of being directly interpretable for the student without reference to the score of any other student, within the context of the definition of the item population. Interpretation of the score without reference to the scores of others is often regarded as a distinctive characteristic of criterion-referenced tests. We shall return to this point in a moment.

Two terms that seem to me to be associated with problems in objectives-based measurement are implicit in this discussion of populations of items and of the role of formal characteristics and of response data in constructing such populations. One is the word "homogeneous." The other is the word "difficulty." The proposition that a population of test items should be homogeneous seems innocuous enough to secure assent from almost everyone. The problem is, of course, to quote Lord and Novick (1968, p. 95): "The question of test homogeneity is one which has been discussed at length in the test theory literature. Unfortunately there is no general agreement as to just what this term should mean and how homogeneity should be measured." They do go on, of course, to suggest a definition that is satisfactory for their purposes; for them, a satisfactory definition is that the components (items) of a homogeneous test are tau equivalent, that is, measure the same trait or "thing." It seems evident that a population of items defined in one of the ways discussed immediately above—that is, defined only with respect to characteristics such as the formal nature of the task(s) and content(s)—is not necessarily homogeneous in the sense of tau-equivalence, which necessarily refers to response characteristics. Thus, in our descriptions of—and allegations about—item populations that are intended to be appropriate for a specific objective a distinction between what might be called "definitional" homogeneity and "response" homogeneity needs to be made.

The term "difficulty" points to additional characteristics of test items that may be relevant in defining populations. Two illustrations will suggest some of the problems buried here. First of all, difficulty does not necessarily imply a normative concept. Certainly it is true that "item difficulty" is often measured normatively, that is, in terms of the proportion of a sample of persons who can answer the item. But consider the Ebel content-standard vocabulary test. For a given student, the difficulty of a particular item might be primarily a function of the number of exposures of that word to the student in the past. If this holds, then it explains rather simply differences in difficulty of two different words for the same student and differences in difficulty of the same word for two different students. Here the nub of the difficulty notion is associated with frequency of experience with the word—the content of the item. Adding pairs of two-digit numbers can be looked at similarly; for a given individual, past practice may determine the difficulty of the item for him, and its difficulty in this sense may not be indicated by normative data. However, we can detect a slightly different view of difficulty—and of its relevance to the definition of populations of items—when we sort out three types of pairs of two-digit numbers: those for which there is no carrying such as 21 and 56, those for which there is a simple carrying such as 38 and 17, and those for which there also is a carrying into the hundreds place such as 68 and 37. We might subdivide or stratify the population of two-digit numbers and thus of our addition items on the basis of this distinction, which postulates differences in difficulty which are associated with process distinctions. If so, we have now attended to the "answers" as well as to the "problems" in defining populations of items.

Further work is needed on the analysis of factors that may enter into the definition of item populations, and thus into the definitions of objectives the achievement of which is indexed by performance on such items. In the real world some item populations appear to be populations by fiat; certainly they lack the definitional neatness (or homogeneity) that is given by stratification of task and/or content dimensions. Specification of several requisite skills for a particular job—skills that appear to be disparate and tangentially related—illustrate such a basis for a population of achievement items. Such an approach may be necessary in some situations; in others such an approach may result from a lack of attention to the analysis of behavior classes and their relations to each other and to the modes of instruction that may be employed.

Let us return now to the specification of types of tests for which a given individual's score is directly interpretable without reference to any other individual's score, that is, without reference to norms. We include the single-item test as one type; the outcome of such a test is simply a statement that the student can or cannot perform this very specific task, and this statement is directly interpretable. As a second type we include the tests made up of a sample of items from a defined population; the score on this type of test, as we have mentioned above, estimates the proportion of the population of items which the student knows or can perform. Given that the population has been defined with some care along the lines we have just discussed, such a proportion (or percentage) is directly interpretable. A third type of test for which the score for a given student is interpretable without reference to the scores of any other student is the test which estimates the rate at which a repetitive task can be performed. Rate tests are familiar in situations such as training students to send and receive Morse code, or training students to typewrite. Measures such as the velocity with which a student can throw a softball probably belong in this class also. A fourth type consists of accuracy measures. A test of accuracy measured by the proportion of trials on which the student is successful, as in a basketball free-throw situation, is similar to the second type, but now the population is a population of trials, rather than of items that are designed to differ among themselves. Dictation tests in stenographic training can be scored in terms of accuracy of reproduction of a given amount and kind of material, and such a score is directly interpretable.

A fifth type of test that yields directly interpretable scores is one that satisfies the restrictions of a Guttman scale. If a set of items satisfies a Guttman scale, the total score on the test, obtained by adding the number of correct responses, describes precisely which items the student passed and which he failed. A trivial example might be these three items: (1)

Add 2 and 2; (2) Solve for x when $5x^2 = 20$; (3) Give the first derivative with respect to x of the function $y = 10x^3$. We would generally expect that any student who answers the third item correctly would also answer the first two items correctly, and that any student who misses the third item but answers the second item correctly would also answer the first item correctly. When this holds, the three items yield only four different response patterns out of the eight possible response patterns. Although ordinarily a total score of 1 for three items might arise in three different ways (answering any one of the items correctly), in our three-item Guttman scale a total score of 1 means that the student answered the first item correctly but no others. Similarly, a total score of 2 means that he answered the first and second items correctly and missed the third. It seems fair to say that Guttman scales of achievement items are difficult to create in practice, except when one uses only three or four items drawn from widely separated levels of achievement, and such restrictions make the test relatively trivial. However, it may be possible to use the Guttman theory in connection with mastery tests for various levels of achievement by using the mastery-nonmastery coding as the items that should scale. If the learning network is properly analyzed into pre-requisites and levels and the mastery tests are properly constructed, then the set of mastery codes should clearly scale in this Guttman sense. Further, it seems likely that shortened versions of each of these mastery tests could be used as a placement test by the simple device of using the mastery-nonmastery code scoring and noting the pattern of these codes for the student. The "perfect" scale patterns are simple to interpret. Additional placement rules would be needed for the hopefully few students who secure non-perfect scale patterns.

I am beginning to feel that a mastery test should be characterized by zero item covariances within a population of tutored and within a population of non-tutored students (but not, of course, within both populations merged). The argument is not immediately relevant, though we may wish to consider it later. In other words, I am beginning to believe that a mastery test should have a zero coefficient alpha within each of these two populations. This is merely to say that the functional efficiency of a mastery test should be concentrated on distinguishing the tutored from the untutored, and not on distinguishing among the tutored or distinguishing among the untutored. If we now consider more than one such mastery test, each appropriate for a different level of achievement, we might expect that this set of tests administered to a random sample of students would yield a high coefficient alpha for the aggregate score or total score over all tests, and that a placement test built from this set of mastery tests would also have a total score of high reliability in this sense. This would not be contradictory evidence. Instead, we are beginning here to make sharper distinctions about item structure within and between populations, a distinction that moves away from the usual single population referent.

I would like to make one more point about mastery tests. It is not unreasonable to regard the outcome of a mastery test as a sign. Let me be stronger. If my notion about the item structure of a properly developed mastery test is correct, then it is quite reasonable to regard the outcome of a mastery test not as a score but only as a sign. If this is so, then the validation of a mastery test poses the problems associated with the validation of a sign. The early article by Meehl and Rosen (1955) should be reread at this point to remind ourselves of the importance of the base rate in determining sign validity. This, in turn, should raise for us the question of what are the proper validity criterion data for mastery tests, and the question of whether or not such data are characterized by naturally meaningful base rates. The problem of developing a statement of sign validity for the mastery test is in part a problem of identifying the appropriate criterion data.

REFERENCES

Bormuth, J.R. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

Ebel, R. Content-standard test scores. *Educational and Psychological Measurement*, 1962, 22, 15–25.

Harris, M.L., & Harris, C.W. *A structure of concept attainment abilities*. Madison: Wisconsin Research and Development Center for Cognitive Learning, 1973.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project*. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hively, W., Patterson, H.L., & Page, S. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, 5, 275–290.

Lord, F.M., & Novick, M.R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Meehl, P.E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychological Bulletin*, 1955, 52, 194–216.

SECTION II

THUS SPAKE PSYCHOMETRIKA . . .

W. James Popham
University of California, Los Angeles

As one of the individuals posing a series of practical measurement problems in the initial set of essays in this volume, I was somewhat distressed that all my concerns were not resolved by the following papers prepared by my psychometrically sophisticated colleagues. I suspect that most of the writers who attempted to establish the problem foci for this monograph were less than completely satisfied. We all wanted nice, tidy resolutions for our criterion-referenced measurement dilemmas. What we receive in the next four essays are answers which in some cases are directly responsive to our questions, in other cases are only indirectly responsive to those questions, and in other cases are responsive to questions unasked. Responsive or not, however, the essays by Harris, Davis and Diamond, Novick and Lewis, and Keesling are full of intriguing ideas which are both timely and thought-provoking. The reader will have to judge whether the four "solution" essays do, indeed, significantly advance our thinking regarding the thorny criterion-referenced measurement problems which face us. These brief introductory remarks will hopefully alert the reader to issues in each of the papers which, to me, appeared salient.

Chester Harris, one of the original architects who conceptualized this monograph, tackled a number of important issues relating to mastery tests. That Harris chose to work with the concept of mastery tests, rather than address the task of defining a "criterion-referenced test," is unfortunate since his keen intellect could have added much insight to this admittedly muddy arena. Among the points Harris makes which I thought warranted particular attention was his analysis of the role of response data in the development of a mastery test. He properly reminds us that the primary use of response data is to provide an estimate not of the suitability of particular test items but of the adequacy of the test plan and the procedures used to generate the items. In a related vein, Harris' distinction between *conceptual homogeneity* and *response homogeneity* is useful in discerning the degree to which a particular test plan does, indeed, generate responses which are homogeneous.

Harris sees utility in distinguishing among members of a learner population based upon *level of instruction*. He sees the possibility of describing an ideal item structure (in which the student who correctly answers one of a set of items would be expected to answer all other items in the set correctly) and then modifying this structure to include different populations identified on the basis of level of instruction. Employing such an approach, Harris continues, "It is possible to make a connection with more conventional test theory."

While this may be true, I think that those working with criterion-referenced measurement should be more than mildly wary when psychometricians somehow manage to massage a problem until it is amenable to the tried and true tactics of traditional test theory. Scholars such as Harris, who were nurtured by and in turn contributed to conventional test theory, may be disposed to view even new measurement problems with their well-honed traditionalist eyeglasses. Fortunately, Professor Harris rarely succumbs to this temptation, in his essay.

Of particular interest was his support for Kriewall's use of Wald's sequential procedure in determining which individuals could be classified as "masters" and "non-masters" on fixed-length tests. Harris' recommendations can be contrasted with (1) the suggestion of Davis and Diamond, (2) the approach proposed by Novick and Lewis, and (3) the scheme suggested by Millman (well described in the Novick and Lewis essay).

There are a number of pithy observations in the Harris paper from which I found myself profiting immensely after a second and third reading. The reader of the Harris essay will have to bring his wits with him; the author obviously did.

The paper by Davis and Diamond deals with a wider range of problems and offers answers which reflect generally accepted practices among measurement specialists. The essay is initiated with a helpful analysis of high points of the early testing movement in this country, leading up to a situation in which criterion-referenced measurement strategies were welcomed. Davis and Diamond then deal with the concept of a criterion-referenced test and, having examined recent conceptualizations, apparently conclude that it can consist of homogeneous items (when used as a diagnostic instrument) or heterogeneous items (when used as a survey instrument). Since their acceptance of heterogeneous items in a criterion-referenced test runs counter to the way many people conceive of criterion-referenced measures, the reader will want to recall this distinction in dealing with the essay by Davis and Diamond.

A number of suggestions are offered for preparing test items for criterion-referenced tests. The attentive reader will note few differences between these suggestions and those that might be proferred for norm-referenced item-writers. Of particular interest was the authors' consideration of four basic principles (used for 25–30 years) as they are related to the selection of items for criterion-referenced tests. As indicated previously, the discussion of test length in the Davis and Diamond paper can be compared with that of several other writers, two of whom offer recommendations in this volume.

For Novick and Lewis the question of optimal test length is not an aside but the chief thrust of their paper. By dealing more completely with this problem, they mount a convincing case that a Bayesian approach to this matter can yield a defensible approach to the determination of minimum acceptable test length.

After dispensing with Millman's non-Bayesian strategy for determining test length, Novick and Lewis set forth a well-explicated design, complete with tables and examples, for determining test length. The focus of the paper is on measurement of student progress in connection with programs of individually prescribed instruction, for in such programs it is imperative to determine whether the student has mastered certain skills prior to permitting advancement to subsequent phases of the program. The paper is concluded by a series of summary observations, several of which are particularly provocative. While the non-quantitatively oriented reader will find the Novick and Lewis paper sticky reading at points, their paper surely represents the most complete published treatment to date of how to resolve the criterion-referenced test length question.

The paper by Keesling cannot be characterized as "easy reading." Keesling offers analytical methods, sometimes highly technical, for dealing with the empirical validation of criterion-referenced measures. This paper will be of particular interest to those who have pondered questions associated with learning hierarchies, for Keesling offers a number of intriguing insights regarding the treatment of tests for such hierarchies.

In overview, then, the second section of this monograph consists of four papers prepared by highly competent measurement methodologists. That the foci they selected seem sometimes removed from the concerns raised by the writers in section one will undoubtedly be of less concern to most readers than to the section one writers. Setting responsiveness considerations aside, however, we find a range of important and compelling observations in the four essays to follow.

SOME TECHNICAL CHARACTERISTICS OF MASTERY TESTS

Chester W. Harris
University of California, Santa Barbara

I shall begin this discussion of technical characteristics of mastery tests by establishing what I think of as the *context* and the *function* of a mastery test. Throughout the paper I regard a test as a systematic procedure for observing and describing a student's behavior. I shall then use this framework as a set of restrictions from which I shall attempt to deduce desirable technical characteristics of a mastery test. These deduced desirable technical characteristics then pose problems of finding ways to index and possibly estimate these characteristics so as to give us information about a particular test. This is a conventional approach to the discussion of any type of achievement test; typically one considers the purpose and nature of the instruction, the kinds of decisions to be made about students and the relation of these decisions to instructional alternatives, and the amount and quality of the evidence that is needed to carry out these decisions. I am here characterizing an achievement test very generally, and I am implying that a mastery test is an achievement test in this sense.

A mastery test is, in my view, necessarily designed to serve the function of determining whether or not a particular student has acquired the behavior or behaviors which a particular instructional program is intended to develop in him. It is a test designed to provide a decision of "mastery" or "non-mastery" for a given student.

It is obvious that this function is not the only one that an achievement test might serve. For example, a diagnostic reading test may be constructed so that particular symptoms or syndromes of perceptual deficiencies can be identified, and then on the basis of the presence or absence of such indicators, instructional alternatives may be selected for the student. One can quibble, of course, and say that these also are mastery decisions—i.e., Has he or has he not mastered this perceptual skill?—but the diagnostic function seems to me to require elements, such as an assumed symptom-treatment network (or theory), which are not necessary for the decision about mastery. It also is obvious that an achievement test may play a role in the study of instruction and be diagnostic in a second sense; here the attempt is to characterize instruction in terms of differentiated outcomes. It also is obvious that by declaring a battery of tests to be a test, we may have "one" test serving several functions.

It may be that the instructional programs for which mastery tests are appropriate consist primarily of instruction in "closed" subject matters, skill-type performances, and the like. There is much discussion of issues of this sort and consequently of the question of the importance of mastery tests. One argument or position rests on the truism that no individual ever "masters" a subject matter or an art, and concludes that a mastery test is a contradiction in terms. Another position regards the evidence of "mastery" as obvious (e.g., one hundred percent performance on a test) and therefore finds the notion of a mastery test to be trivial. (I am distinguishing here a *concept of mastery* from a *mastery score or index*, and I shall elaborate these ideas later.) Apparently the first argument finds the concept of mastery undefinable and thus unavailable for use, whereas the second mistakes the mastery score or index for the concept of mastery.

I would like to cut through all this by taking the position that in at least some instructional situations it is important to have available a means of making a "mastery"–"non-mastery" decision for a given student. This is the proper function of a mastery test. I also wish to require that a mastery test or set of mastery tests function within a particular instructional program consisting of one or more units of instruction, each characterized by one or more specific objectives. This requirement indicates that I view a mastery test not as a "general" achievement test with respect to its subject matter, but instead as an objectives-based test that is restricted to one or more specific objectives. The test that is designed to serve this function is the subject of this paper.

CONCEPT OF MASTERY

Specifying the concept of mastery is part of the problem of developing the instructional program as well as part of the problem of developing the mastery test. Let me outline some steps that provide a path to the specification of the concept of mastery. The objectives of the relevant portion of the instructional program should be specified with respect to the critical behaviors, the materials or content with which these behaviors deal, and the conditions or situations in which the critical behaviors are expected to be displayed. A single behavior-content-situation intersection in the statement of objectives defines an achievement for which a mastery test can be constructed. Next, it is necessary to define or select one or more types of test item or task that will provide what is judged to be the critical evidence of the achievement of the specified objective. If techniques such as those described by Hively and his associates (1973) are appropriate, they can be employed to define a population of such items or tasks either by a complete listing of the tasks or by specifying a mechanism for generating such items or tasks. This is the procedure of translating a statement of one or more behavioral objectives into specific test tasks, the performance on which provides a basis for an inference about mastery or non-mastery of the objective. Given such a population of items, it then should be possible to state the smallest proportion of this population of items that should be responded to correctly by the student if he is to be considered as having mastered the instruction. This judgment yields a concept of mastery that is expressed as a minimum level of performance on a defined population of items or tasks. Note that this does not solve the problem of what test score based on a sample of items should be taken as a mastery score or index. This is a separate problem that is considered later.

MODELS OF ITEM STRUCTURE

The homogeneity or lack of homogeneity of the population of items should be commented on. I wish, in the process, to distinguish between *conceptual homogeneity* and *response homogeneity*. Consider Glaser and Nitko's (1971, p. 655) illustration of the domain of 111,000 items consisting of all 3-, 4-, and 5-addend column addition tasks with the restriction that each addend be a single-digit integer from 0 through 9. The one thousand 3-addend problems form a subset that may be conceptually different from the 4-addend, and the 4-addend different from the 5-addend problems; consequently the domain might be regarded as divisible into at least three distinct populations of tasks. Within the 3-addend problems, however, further distinctions might be made. For example, problems for which the answer is a 1-digit number may differ from those for which the answer is a 2-digit number, and so the 3-addend problems might be sorted into two populations on this basis. Further analysis could yield further subdivisions and smaller populations.

At some point in the analysis it should be possible to state that for students for whom the particular instruction is appropriate, such and such a set of tasks constitutes a conceptually homogeneous population in the sense that the individual tasks are interchangeable *for the purpose of determining mastery*. It is desirable to arrive at this point, since the concept of mastery is stated in terms of a student's performance on a population of items that may be extensive; if so, the only feasible course is to test the student on only some of the tasks. The question of which tasks are to be used arises. If the tasks are interchangeable for the purpose of determining mastery, then the proper strategy is to test a given student on a *random* sample of the tasks, since this yields an observed proportion correct (of the sample of items or problems) which is an unbiased estimate of the proportion of the population of tasks that he could solve had he been given them all. The term *random* is critical; it implies that the items are interchangeable for this purpose.

The concept of mastery should be specific to the instructional program and to the population of students for whom that program is appropriate. To use the Glaser-Nitko example again, 2-addend column addition problems might be an appropriate conceptually homogeneous set of items from which to create mastery tests for younger students who are beginning instruction in addition. However, at a later stage of instruction, 2-, 3-, and 4-addend problems might all be merged into a single population that would be regarded as conceptually homogeneous for students who had completed this later instruction. In general, this homogeneity should be conceptualized with respect to a particular population of students; namely, those who enter the instruction with the necessary prior achievements and complete the instruction under a reasonable schedule. This point bears on the question of differences in difficulty of various items.

It is important to distinguish between the difficulty of an item for a given student, which presumably is strictly a function of his experiences and his previous instruction (both formal and informal), and the normative concept of difficulty, which depends as well on who makes up the group of students whose responses provide the estimate of difficulty. For example, it is reasonable to believe that adding 9 and 8 may be more difficult than adding 2 and 4 for a given student at an early stage of his instruction. However, if the instruction properly adjusts the exposures and the practice, there should come a time during the instruction when these two items will be of approximately equal difficulty for that student. I say "approximately" because there may still be an infinitesimal difference in latency in free responding to the two items which suggests that all differences have not been eliminated. However, this approximately equal difficulty, which means simply that the student can readily provide the answer to either problem on request, is sufficient to regard the two items as interchangeable for the purpose of determining mastery. It is in this sense and at this stage of the instruction that conceptually homogeneous items are of equal difficulty for a given student.

Macready and Merwin (1973) use Hively's ideas about a domain-referenced testing system in considering the relation among items within an item form, which is a population of items in the sense I have been using those terms. For Macready and Merwin, the ideal for a diagnostic test would be that whenever a student gets one item within an item form (population) right, he would then get all items within the item form correct. I assume that this ideal might also characterize a mastery test. The implication is that all items are of equal difficulty for that student and the conditional probability of his passing a randomly chosen item, given that he has passed another item in the item form, is unity. Now if there is a population of such students, then their item response data generate a matrix consisting of all 1's, where 1 designates a correct response; and for any sample of students drawn randomly from this population, the items are of equal normative difficulty. Further, the items are characterized by response homogeneity as evidenced by this conditional probability of unity even though the *correlation* between any pair of items is not defined because of zero variance for each item.

If we add a population of students who because of lack of instruction can respond only randomly to these items, with the probability of success on any item close to zero, we expand the matrix by adding rows consisting primarily of zeroes for these untutored students. Sampling from both student populations in effect selects only some of the rows of this matrix to examine. If these sample data for the various items are used to estimate item difficulty, we would expect that the items would have very similar normative difficulties, but that the absolute value of the difficulty level so estimated would now lie somewhere between zero and unity, depending upon the proportion of students entering the sample from the two populations. Again, response homogeneity would be evident in the sample data, now with conditional probabilities of answering a randomly chosen item correctly, given that another item has been answered correctly, close to unity for the students drawn from the one population, and close to zero for those drawn from the other. The big difference would be that now the pairs of items would be characterized by determinate, and almost perfect, correlations, and the correlation of an item with total score (as in a biserial or point-biserial coefficient) would be large.

There probably are subject matters for which this model of item functioning is quite reasonable. The use of a transit in surveying, knowledge of English equivalents of certain Arabic nouns, knowledge of rules for forming derivatives of various types of functions, etc., illustrate these. Generally, those persons who have not been instructed should fail the item(s); those who have been instructed should be able to perform successfully. To the extent that successful performance on the items is a function of instruction, the normative difficulties of the items reflect only the levels of instruction represented in the sample of students, and the item correlations (or functions of them, such as Loevinger's index of homogeneity which Macready and Merwin used in their study) reflect only differences in these levels of instruction. The simplest model postulates only two populations of students: the well instructed and the untutored. This model ignores a potential population of students who are "in process" and who, on a relevant mastery test, are likely to secure a range of scores rather than be bunched toward the top or toward the bottom. However, the model can be modified to include several—possibly indefinitely many—populations of students who differ in level of instruction.

Such a model is most tractable if we have, for each level of instruction, a constant conditional probability of answering an item correctly, given that the student has answered another item correctly. In the two-population case discussed above, we set this conditional probability at unity for the tutored population and at zero for the untutored. If—as seems likely—there are populations of students who can do "most" of the items but not all, then it would be desirable to have a constant conditional probability slightly less than unity for such a population; a consequence is that we still have for this new population of students interchangeability of the items for the purpose of estimating from a sample of items the proportion of the population of items which can be responded to correctly, i.e., estimating the numerical value of this conditional probability. We can make a similar statement about a population of students who can do only "a few" of the items. For this population a constant conditional probability slightly greater than zero would be associated with interchangeability of items for the purpose of making this estimate. It should be pointed out that a model of item structure which specifies constant conditional probabilities within student populations that differ in level of instruction ("ability") is a model of local independence in "classic" test theory. Thus, beginning with Macready and Merwin's "ideal" item structure and then modifying it to include many populations differentiated on the basis of level of instruction, it is possible to make a connection with more conventional test theory.

One distinction I am making that may be of importance is indicated by my using "level of instruction" to describe student populations. Conventionally, the term "ability" is used, and then "ability" is assumed to be continuous and (possibly) normally distributed. Such a formulation has proved its utility in a number of areas. However, here it may be inappropriate to assume that level of instruction is a continuous variable for the purpose of mastery testing. Instead, the transition from one level of instruction for a student—and thus for a population of similar students treated similarly—may be a quantum jump rather than a continuous accretion. Something like a familiar all-or-none model may be the proper conception. This, of course, is debatable. Even if continuity is not an issue, however, level of instruction clearly is a variable that does not have a "naturally occurring" distribution of values that remains stable; instead, by altering instruction it is possible (at least in theory) to alter the relative frequencies of students at the various possible levels of instruction. If this view is correct, then making a general assumption about the shape of the distribution of level of instruction is rather hazardous. Instead, one is prompted, in any empirical study, to control this distribution by controlling the character of the instruction. One can then examine the empirical item structure for samples drawn from the selected levels of instruction to determine the extent to which constant conditional probabilities are well approximated at the various levels being studied. Again, the simplest model is the two-level model consisting of instructed and non-instructed students, and one might expect it to be the easiest one to deal with empirically.

ROLE OF RESPONSE DATA

There appear to be two rather different points of view about the role of response data in the development of a mastery test. In the section above I have outlined the model of item structure that seems to me to be called for when one wishes to use a sample of items as a basis for estimating performance on a population of items. Such a procedure, as Kriewall (1969) has argued, gives a test score (a proportion or a percentage) that is directly interpretable without reference to the performance of any other student. The score is, in this sense, criterion-referenced rather than norm-referenced. If we wish to preserve this feature, it is essential that the particular sample of items making up a particular mastery test be drawn at random from the population of items and be retained in the test, regardless of their response characteristics. I have tried to emphasize that for the resulting mastery test to be a reasonable one, great care must be taken in the specification of the objectives of the instruction and of the character of the instruction so that a conceptually homogeneous population of test items or tasks is defined, and so that instructional procedures that will insure interchangeability of the items for the purpose of determining mastery are provided. If this is a sound position, what then is the role of response data in the construction and/or study of a mastery test?

I conclude that the construction of the particular test should proceed without attention to response data. What I am saying is that it is not the particular sample of items making up a particular test that is of great concern. Instead, it is the plan that permits the development of randomly equivalent alternate forms of the mastery test that is critical. Response data gathered for one of these forms ought to be used, not to modify or exclude items within that one form, but to secure estimates of how well the plan has been conceived and operationalized. Thus, if response data for a particular form of the test indicate that certain items do not function as anticipated, then the search should be for what went wrong in the process of defining the objectives of instruction, conceptualizing an appropriate population of items, and deducing the instructional procedures that are required to bring students to the desired level of performance. This position views the "bad" item as a symptom. Rather than arbitrarily excluding the "bad" item, we should alter the system so that the "bad" item either will not be included in the population of potential items because of a new conception, or will function differently because ideas about the relation between the instruction and the evidence of achievement have been modified.

The study of a particular mastery test that consists of a random sample of items or tasks from the specified item population is critically dependent upon the population(s) of students sampled. This point has been emphasized earlier in the discussion of student populations classified in terms of level of instruction. With the simplest model, one wants to sample a population of instructed or tutored students and a sample of non-instructed students. If the system for developing the mastery test is functioning properly, one predicts approximately equal item difficulties within each of the two populations of students, but difficulties that differ systematically between the two populations. One also predicts response homogeneity for items within each population of students, but recognizes that phi coefficients close to zero for item pairs within a population do not necessarily indicate lack of such homogeneity.

Two kinds of studies appear to be especially important. One attempts to answer the question: How well does the test sort students into two groups? The other attacks the question: How well does the test sort students into the correct two groups? These correspond at least roughly to questions of reliability and of validity in more conventional contexts. I shall turn to them shortly, but first I wish to discuss very briefly the role of the mastery score or index.

THE MASTERY SCORE OR INDEX

Recall that I defined above a *concept of mastery* and operationalized the term as a proportion of a population of items that a properly instructed student should be able to answer correctly. If this proportion is generally not observable but instead must be estimated, we face the problem of determining for a given student at a given testing the mastery score or index that must be reached on that test for us to regard the student as a "master," i.e., to decide he should be advanced to new instruction. We may also wish to set a second cutting score such that a student who scores at or below this point is regarded as a "non-master," with any student who scores in between being "in limbo." This sorting into three categories on the basis of a fixed-length test is feasible, and the necessary development is available in Wald (1947, Chapter 5). Kriewall used the Wald sequential procedure as a basis for a proposed computer testing operation in which an individual student would receive and respond to as many items as needed in order to make the "mastery"–"non-mastery" decision for him. However, the adaptation of the Wald procedure to a fixed-length mastery test is actually provided for by Wald's discussion of "Observations Taken in Groups" (1947, pp. 101–103).

In order to illustrate the adaptation of the Wald procedure to a fixed-length mastery test, consider the case of an 8-item test when one wishes to declare the student a master if he could pass 90% or more, and declare him a non-master if he could pass only 70% or less of the population of items. Associated with these decisions is a willingness to make an incorrect decision of "master" 25% of the time and an incorrect decision of "non-master" 25% of the time also. These values may be substituted in the Wald equations to determine that the mastery score or index should be 100% of the 8 items, and the non-mastery index 62% of the 8 items. With this fixed-length test, a student who scores between 62% and 100% cannot be assigned to either category. Giving him a second (but different) 8-item test and scoring the two together would permit advancing him to new instruction if he scored 88% or more on the 16 items, or declaring him a non-master if he scored 75% or less. For 16 items there is still a middle region of no decision.

The procedure illustrated here may be of interest as a method of defining the mastery score or index. It is relatively simple to implement for fixed-length tests, and it has the important feature of incorporating specified risks of the two types of error. Further, it tends to show that very short tests often do not lead to informed decisions. It should be preferred to the use of the tables presented by Millman (1973) since Millman's work solves the "wrong" problem. For further discussion of this problem of how many items are needed and what the mastery score or index should be for that number of items, one should refer to the paper by Novick and Lewis in this monograph, which has the interesting feature of incorporating collateral information in a Bayesian framework.

AN INDEX OF EFFICIENCY

Let us now turn to the question of what I shall call the efficiency of a mastery test. I am aware that this term is used in connection with classic reliability theory, but at the outset I do not intend to imply that the term *efficiency* means all that reliability means.

I have argued that a necessary characteristic of a mastery test is that it sorts students into two categories. If in addition the test is valid, it will tend to sort them into the *correct* two categories: that is, into the categories determined by the criterion data. In the absence of such criterion data, it may be informative simply to examine how well the test sorts defined samples of students into categories and possibly to measure its efficiency in this sense. It is important to point out that we are not breaking a new path here, since as early as 1936 Richardson (1936) considered this problem of a "criterion of two categories" using scores on a total test cut at various points as the "criterion." His work relates the difficulty of a test element to the prediction of a two-category criterion, employing certain distributional assumptions. We shall attempt a similar development making somewhat different assumptions.

Let us assume that a mastery test consists of k items and that a total score on the test is derived by summing the number of correct items, which gives zero and k as the limits for any score. Let us also think of these items as ones for which the student produces a response, rather than chooses a given alternative. With total scores ranging possibly from zero through k, there are k different possible separations into two groups on the basis of total test score. For example, students who score k may be sorted into one group and all others into the other group; students who score at least k-1 may be sorted into one group and all others into the other group; etc. Thus, there can be k different sorts. For any sort, let us develop an index that is suggested by Fisher's linear discriminant function for two groups (Fisher, 1936). The discussion by Tatsuoka (1971, Chapter 6) is quite helpful since he shows canonical correlation equivalents of discriminant functions.

By a "sort" we mean that the sample has been sorted into two groups on the basis of some cutting point on the total score for the k items. We can then, following Fisher, develop two k-by-k matrices, B and W. (See Tatsuoka, pp. 158–159). The matrix W is the pooled within groups sum of squares and cross products of the item responses. The matrix B equals T - W, where T is the sum of squares and cross products of the item responses, ignoring the separation into two groups. Then given the group membership, the Fisher discriminant function is

$$\frac{v'\,B\,v}{v'\,W\,v} = \lambda\,,$$

where v is a column vector of weights chosen to maximize λ. Instead of using these weights, let us use an a priori vector of equal weights, 1, and form the function

$$\frac{1'\,B\,1}{1'\,W\,1} = \lambda_c$$

which is a special case (equal weights) corresponding to using the total score (sum of the item scores) to discriminate the two groups. Generally λ_c is less than λ.

Now λ_c turns out to be a function of the sums of squares associated with the two-group analysis of variance. It is

$$\lambda_c = \frac{SS_b}{SS_w}\,,$$

where SS_b and SS_w refer to the analysis of the total scores on the k items for the two groups. We also know that in general the Fisher discriminant function can be related to a canonical correlation between the given variables (items) and a dummy variable indicating group membership. In general, if μ^2 is the squared canonical correlation, then

$$\lambda = \frac{\mu^2}{1 - \mu^2}\,,$$

and

$$\mu^2 = \frac{\lambda}{1 + \lambda}\,.$$

An analogous treatment of λ_c yields

$$\mu_c^2 = \frac{\lambda_c}{1 + \lambda_c}$$

or

$$\mu_c^2 = \frac{SS_b}{SS_b + SS_w}\,.$$

This coefficient is equivalent to the squared Pearson product-moment correlation between total score on the test and the dummy variable designating the sort. Thus it is the squared point-biserial correlation coefficient.

Sorting into two (non-empty) groups on the basis of total test score necessarily yields a positive value for SS_b and thus a positive value for μ_c^2. The upper limit of μ_c^2 can be $+1$ when $SS_w = 0$; this could occur, for example, when only two different total scores appeared in the sample and were sorted into the obvious two groups. Such a situation would correspond to a perfect *phi* coefficient.

The coefficient μ_c^2 for a given sort based on total score measures the extent to which the sum of the k item scores (0, 1 scores) can discriminate the two groups defined by the sort. It is a measure of efficiency in this sense and has two features that make it an analog of a classic reliability coefficient. One is that it can be conceived as the ratio of true score variance to observed score variance for a particular definition of true score. To achieve this correspondence, assign to each individual in the upper group a true score equal to the mean of the upper group and to each individual in the lower group a true score equal to the mean of the lower group. Then μ_c^2 will be the ratio of the variance of these assigned true scores to the variance of the observed scores. Note that μ_c^2 was defined originally without reference to true score variance and that we have now simply answered the question of how true scores might be conceived to make μ_c^2 an analog of the classic reliability coefficient.

The second feature is that the largest μ_c^2 for a given test is an upper limit to the validity of the mastery test when validity is measured in an analogous fashion. First note that for a k-item test there are k different sorts into two groups based on total score and that there is a value of μ_c^2 associated with each sort. Suppose now we have a dichotomous criterion and use this, rather than total score, to sort students into two groups. If we now measure in a similar fashion the extent to which the sum of the item scores can discriminate the two criterion groups we find that this coefficient cannot exceed the largest μ_c^2. It may of course be substantially smaller. It also is true that if the two criterion groups are not equal in number, the upper limit will be some μ_c^2 less than the maximum and corresponding to a sort into two groups with the same relative frequencies.

It is possible to deduce some generalizations about maximum values of μ_c^2. For example, for symmetric distributions the maximum value of μ_c^2 occurs when the proportion in the upper (or lower) group is close to one-half and decreases as this proportion diverges from one-half. For symmetric distributions of equal range, a rectangular distribution gives a larger maximum than does a normal distribution. It is intuitively evident that a U-shaped distribution has a larger maximum coefficient than does a rectangular distribution of the same range. Interestingly, a rectangular distribution of small range has a larger maximum coefficient than does a rectangular distribution of large range, though the difference may be small.

It may be worthwhile to emphasize the point that this index, as a descriptive measure, can be developed without conventional distributional assumptions for the item scores or for the total score. Such assumptions are needed for testing hypotheses, but the tests themselves may be reasonably robust and thus not sensitive to violations of the distributional assumptions. Further, in most uses of an index like this, testing a null hypothesis probably is not of any great interest. It would be desirable, of course, to have a confidence interval for the statistic, but at the moment I do not see a solution to this problem that does not involve possibly restrictive distributional assumptions. When a mastery score or mastery index has been specified for a test of a particular length, then this mastery score can be taken as the cutting point and the coefficient computed for that particular sort into two groups. This relates the coefficient I have described to the procedure for determining whether or not a student has reached the level specified by the concept of mastery that was adopted. If both a mastery index and a non-mastery index are employed, as in the adaptation of the Wald scheme, two cutting points are identified. One might, of course, compute the two coefficients in this situation. An approach that appears to be better would be to perform the three-group dispersion analysis rather than the simpler two-group Fisher analysis, but compute only the coefficient associated with the a priori vector of equal weights. This suggestion needs study and trial with actual data.

APPROACHES TO VALIDITY

For a mastery test, the ultimate validity question is the question of the extent to which the test sorts students into the correct two categories. Given an appropriate criterion, it is possible to develop the two-by-two table that results from classifying students as "true masters" or "true non-masters" on the basis of the criterion data and simultaneously classifying them as "indicated masters" or "indicated non-masters" on the basis of the mastery test. An appropriate interpretation of these data provides a validity statement for the test. The path leading to such a result may be a rather long one. Some of the points along this path can be conceptualized as marking necessary steps for completing the journey.

The relation of the mastery test to the instructional program—the context of the test—is critical. Thus the test construction process does not begin with an item bank left over from another project or with a survey of the latest achievement tests. It begins with a careful specification of what is to be learned and how this is to be learned. From here the test developers can move to the specification of the items or tasks that are judged to provide the critical evidence of the achievement. A concept of mastery may then be established. This judgment should be made in relation to such features of the instruction as the amount and distribution of practice, the kinds of feedback, etc. The concept of mastery should be tentative, since the level at which it is set may be a factor in the immediate and/or subsequent achievement of students. Some of the necessary evidence concerning this can be gathered fairly quickly; some may require a much longer study. The selection of items to make up a particular mastery test should proceed according to a sampling rule and not be determined by personal views of particular items or by study of response data. The test itself is only one of many such possible tests, and the study of a particular mastery test should be designed to throw light on the system for developing alternate forms of the test. When such study indicates that the system functions much as anticipated and that tests so constructed sort students effectively into two or possibly three categories, the study of the relation of test decision to criterion data becomes a culminating step.

Several types of criterion data can be identified. In speaking of level of instruction as a variable and in suggesting that this variable be controlled in empirical studies, I have implied that the prior instructional history of the student may be relevant criterion data. Ozenne (1972) has presented ANOVA paradigms for studying a test in terms of its sensitivity to instruction; he discusses a pre-post model, using the same sample of students before and after instruction, and a model using samples of two independent groups, one of which was instructed and the other not. This second model can be related to the index of efficiency described in the preceding section. For both of Ozenne's models the items making up a test should be a random factor, rather than fixed. Ozenne also assumes that the items or measures are "comparable"; here this probably means *tau* equivalent in the classic sense. Items of a mastery test drawn randomly from a population of items that is constructed in the manner that I have outlined earlier should fit this model reasonably well. In the pre-post model Ozenne does not sort out effects due to items or to the interaction of items with subjects and/or occasions, but instead pools these as the source of error of measurement. He then estimates the variance due to occasion (from the occasions and the occasion-by-subjects mean squares, since subjects is a random effect) and compares it with the estimate of the variance due to occasion plus the variance due to measurement error. This is taken as an index of sensitivity to instruction. The two independent group model makes a similar estimate using a nested design. Ozenne's work provides two approaches to the use of students' instructional history as a criterion variable in studying the validity of a mastery test.

A second type of criterion that may be employed in the study of the validity of a mastery test is performance on a transfer task. Such a task may, of course, demand proficiency on more than one component of achievement, and thus may be a criterion for a set of mastery tests. Field tasks that demand the solution of a problem in a non-classroom setting, applications that demand the integration of processes, and extrapolations that demand the transfer of principles are possible illustrations of transfer tasks. If the transfer task is a relevant criterion for only a single mastery test, then the problem of summarizing the data can be formulated as a problem of interpreting the two-by-two table described at the beginning of this section. If, however, the transfer task is a relevant criterion for a set of mastery tests, it is necessary to consider alternatives in the representation of the mastery test data. Thus, a simple composite of mastery test raw scores might be related to the criterion. However, a composite necessarily weights the various test scores in some fashion, and this weighting can affect the magnitude of the relationship with the criterion. Use of a composite, then, poses the problem of selecting a weighting scheme for the various mastery tests. Elsewhere in this monograph I have suggested that several mastery tests might be combined in a different fashion for a purpose such as studying validity. One would score each mastery test as "master" or "non-master" and then examine the pattern of these "profiles" for students at different levels of the criterion. If the achievement indexed by each mastery test contributes equally and independently to performance on the criterion, then this should be reflected in a substantial relationship between the number of "mastery" scores for a student and his criterion score; in this situation the different ways of securing a given number of mastery scores would not distinguish among criterion scores. If instead these achievements form a well-defined hierarchy leading to the criterion behavior, then this situation should be reflected in the occurrence of many profiles that are scalable in the Guttman sense and very few profiles that do not scale, along with a substantial relationship between the number of mastery scores and the criterion score. It is possible that the set of mastery tests might scale, but that this scale would not be related to the criterion scale; such a finding would prompt a reexamination of the nature of the criterion.

A third type of criterion for a mastery test is degree of subsequent success. This is simply a predictive validity notion, in which performance on a mastery test is related to subsequent success in the continuing program of instruction. The criterion may exist in varying degrees of remoteness ranging from the next instructional unit to completion of the school program. Note that this rather conventional approach to validity might also be regarded as a use of the students' instructional history as a criterion, but that it differs from the first type of criterion in that the first type looks back whereas this looks forward. The first postulates that the effect of instruction can be detected in the students' performance on a mastery test; this predictive approach postulates that performance on a mastery test at the conclusion of a unit of instruction is indicative of future performance. Thus these two types differ in an important way.

Undoubtedly other analyses of the problem of assessing validity of a mastery test or set of mastery tests might be made. The analysis given here tends to stress first what might be called a combination of content and construct validity, the evidence for which rests to a considerable extent upon the record of how the test was conceived, related to the instructional program, and developed. I then suggested what seem to me to be three somewhat different types of criterion data that might be employed in relating test performance to criterion. That there may be more than one kind of criterion data emphasizes the point that defining "right" in the study of the extent to which a mastery test sorts students into the "right" two groups is not a trivial problem.

I shall close this section by returning to the two-by-two table and commenting on the problem of summarizing such data. It is obvious that there are many coefficients or statistics that might be used to summarize such a table; two are especially familiar to persons dealing with tests: the *phi* coefficient and the tetrachoric coefficient. It also is evident from the literature that not all "experts" agree on which, if either of these two, should be used. Possibly in a situation like this it would be useful to turn our attention to another statistic that does not belong to the correlation family. This statistic is a conditional probability.

For a table such as the one I have described, two quite different conditional probabilities may be computed. One would be the conditional probability of a student's achieving a satisfactory criterion score (being judged a "true master") given that he scored high on the mastery test (being judged an "indicated master"). This conditional probability can be estimated from appropriate sample data by determining the proportion of those students scoring high on the mastery test who also score high on the criterion. When it is properly estimated, this conditional probability is the likelihood that a student randomly selected from the group of "indicated masters" would be found to belong to the group of "true masters." For example, if the criterion is subsequent performance in an instructional program, then this conditional probability indicates the proportion of those who "pass" the mastery test who will later secure a high criterion score. The difference between this conditional probability and the "base rate" (the proportion of students who do well in the subsequent instructional program, regardless of their mastery test score) can then be related to the *phi* coefficient. A second, and different, conditional probability is given by the proportion of those students scoring high on the criterion who also score high on the mastery test. This coefficient describes the likelihood that a student randomly selected from the group of "true masters" would be found to belong to the group of "indicated masters." An important question is: Which of the two conditional probabilities is of interest? They answer different questions, and therefore should not be confused or assumed to be the same.

Within the context of studying the validity of a mastery test, a use of this second conditional probability may be the following. Let the criterion data be the instructional history of the student, i.e., knowledge of whether or not he was given formal instruction in the materials covered by the mastery test. Then this second conditional probability would describe the likelihood that an instructed student would score high on the mastery test. As such it attends only to the decision of "indicated master" or "indicated non-master," based on total score, rather than examining performance on the items as does an Ozenne paradigm. This "compact" procedure may be useful in certain situations. The point to be emphasized here is that estimating the probability that a student will do well on the mastery test, given that he was previously instructed, may be much more meaningful than estimating the probability that he was instructed, given that he did well on the test. This latter is often thought of as a "postdiction" in contrast to a prediction.

Elsewhere in this monograph Keesling discusses uses of conditional probabilities in much more detail.

SOME CAUTIONS

I wish to mention some cautions that the reader should observe in interpreting what I have said here. The paper represents an attempt to define a mastery test and then to deduce—hopefully with a minimum number of assumptions—desirable technical characteristics of such a test. The paper also suggests ways to index or describe some of these characteristics, but for the most part in only a preliminary and tentative fashion. Thus the discussion is suggestive—and may be useful in that sense—but is not viewed as complete.

The paper in no way tries to wrestle with what appears to be a morass of ideas and statements about "criterion-referenced tests" in the literature. (I used the term "criterion-referenced" only once in the paper, and then to describe a score rather than a test.) This merely indicates that at the present time I do not understand much of what has been written on this topic; and by refraining from adding to this literature I hope to make a contribution by omission rather than commission. In contrast, I feel that the mastery test can be defined reasonably precisely and that principles that should underlie its construction and study can be identified. This was the task to which I addressed myself. It is my hope that similar, but necessarily different, analyses of other types of achievement tests will be undertaken. For example, tests that are designed to provide feedback to students during instruction deserve attention; analysis may indicate that they are distinctive in several ways and that their construction and use should be governed by speeific rather than vague, general rules. By identifying functions to be performed by achievement tests and then bringing our "test theory" to bear on these problems, we may in time be able to present a more detailed map of the achievement territory.

Some of the things I have said in this paper may be slightly unrealistic at present. Let me give one example. Early in the paper I talked about populations of tasks or items that can be defined completely, either by enumerating all of them or by specifying an item generation system. We can readily determine such a population in many instructional situations; however, there probably are subject matters and instructional objectives for which this is now impossible. I have two defenses. One is that I am optimistic and would hope that for some of these situations the study of curriculum and instruction will enable us to move in the direction of specifying objectives in terms of definable populations. The other is that I make no claim that all achievement tests should be mastery tests. The mastery test serves a particular function, in my view; and if this function is not appropriate, then the test is irrelevant. It may be that a major curriculum question for any instructional program is simply the question of the appropriateness of the concept of mastery testing within that program.

Finally, I must acknowledge that some of the ideas here were developed in conversations with Peggy, even though in true male chauvinist fashion I have not added her name as an author.

REFERENCES

Fisher, R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, Vol. III, Part II, pp. 179–188.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational Measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project*. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Kriewall, T.E. Applications of information theory and acceptance sampling principles to management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Macready, G.B., & Merwin, J.C. Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement*, 1973, *33*, 351–360.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205–216.

Ozenne, D.G. Toward an evaluative methodology for criterion-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1972.

Richardson, M.W. The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1936, *1*(2), 33–49.

Tatsuoka, M.M. *Multivariate analysis*. New York: Wiley, 1971.

Wald, A. *Sequential analysis*. New York: Wiley, 1947.

THE PREPARATION OF CRITERION-REFERENCED TESTS

Frederick B. Davis and James J. Diamond
University of Pennsylvania

Criterion-referenced tests fall into the category of measuring instruments called achievement tests. Achievement tests are the oldest and most widely used type of measuring instruments. The Chinese used such tests in selecting civil-service employees thousands of years ago and teachers and tutors since time immemorial have used oral and written examinations to determine how much their pupils have learned of the content covered by their teaching. In 1864, for example, Chadwick (1864) described a *Scale Book* prepared by the Rev. George Fisher which, he said,

> . . . contains the numbers assigned to each degree of proficiency in the various subjects of examination . . . The numerical values for spelling . . . are made to depend upon the percentage of mistakes in writing from dictation sentences from works selected for the purpose, examples of which are contained in the *Scale Book* in order to preserve the same standard of difficulty.

Needless to say, a great deal of time and effort has been devoted to ensuring the adequacy of achievement tests. By the middle of the twentieth century a large body of information had been accumulated regarding the planning and design of achievement tests. A convenient summary of this material was provided by Lindquist and others in *Educational Measurement* (Lindquist, 1951, Chapters 5–13). In Chapters 14 through 18 this book also brought together technical material on the development of achievement tests. At about the same time, Gulliksen (1950) published *Theory of Mental Tests*, which dealt with the statistical bases of test construction in a systematic way. By the middle of the century, therefore, a theory of achievement testing had been created that prescribed in great detail methods for identifying and defining the areas or "domains" of subject-matter content (facts and understandings), skills, attitudes, and feelings to be measured; procedures for preparing tasks or test items to measure the objectives; and statistical techniques for assembling and refining items and expressing test scores in various scales or "metrics" and for providing interpretive information.

During the 1950's and 1960's the migration of hundreds of thousands of rural families from the South and from Puerto Rico to more Northern cities in the United States and the marked increases in the percentage of teen-age boys and girls in attendance at secondary schools focused attention on some inadequacies in traditional curriculum materials and teaching procedures in American schools. Educators and psychologists

120

responded by reexamining the objectives of the schools and ways for formulating these objectives. Task analysis and the detailed specification of sub-objectives for instructional purposes, often in terms of observable and measurable behaviors, were undertaken; systematic instructional programs, or procedures, were devised, tried out, and refined in a revitalized effort to ensure that each pupil would master the essentials of the basic school subjects, especially arithmetic and reading. The goal of mastery in individualized instruction that was emphasized by Morrison (1926) during the 1920's in his Unit-Mastery Plan, by the Dalton Plan (Parkhurst, 1922), and by the Winnetka Plan (Washburne, 1932) was emphasized. In the learning of basic school subjects, the factor to be varied was perceived to be the time required for different pupils to master knowledge and skills—not the degree of competence acquired by different pupils in a predetermined amount of time.

Naturally, techniques for assessing the acquisition of knowledge, skills, and feelings by school children changed as educational objectives became more specific and more diversified. In fact, individualized teaching programs depend for their successful operation on the intertwining of teaching and assessment procedures, whether the latter are formal or informal, oral or written. About 10 years ago Glaser and Klaus (1962) discussed the analysis of instructional objectives and Glaser (1963) emphasized the importance of meshing instructional procedures with evaluation instruments especially designed to guide individualized instruction. In the educational climate of the mid-1960's this basic idea was extraordinarily appealing because most teachers were not finding reports of standardized testing programs sufficiently informative to be helpful in guiding their teaching and because many educators judged widely used standardized tests to be inappropriate and even unfair for testing pupils, especially pupils from underprivileged and minority groups.

CRITERION-REFERENCED TESTS

The scene was thus set for a deemphasis of standardized survey tests of achievement with their national norms, converted-score scales, and assorted statistical techniques for interpreting scores. Educators sought tests keyed to specific, observable, realistic objectives. Since a variable that a set of items (or tests) is intended to measure has long been called a "criterion" by psychometricians, a test made up of items carefully constructed to measure an individual's performance on a task or tasks defined as the objectives of instruction was called a "criterion-referenced" test. Nitko (1970) stated this plainly: "A criterion-referenced test is one that is deliberately constructed to give scores that tell what kinds of behaviors individuals with those scores can demonstrate (p. 38)." This definition clearly implies that criterion-referenced tests are intended to be used as diagnostic instruments for identifying behaviors that examinees can or cannot perform. It also implies that great care must be exercised in preparing the outline or plan for a criterion-referenced test to make sure that a representative sample of all of the behaviors that the objectives of instruction call for is measured by the items.[1] Yet the definition does not preclude the inclusion in a single criterion-referenced test of items that measure performance on widely different kinds of behaviors that may be only loosely correlated with one another even if they are measured with perfect reliability. Let us define a set of items that comprise a test as homogeneous if all items measure the same variable, plus random error; let us define a set of items as heterogeneous if all items measure entirely uncorrelated variables, plus random error. These two limiting cases are not found in practice so we shall use the terms "homogeneous" and "heterogeneous" to characterize sets of items that approach one or the other of the limits.

Unless all of the items in a test measure exactly the same variable, or variables for which true scores are highly correlated (say, .90 or greater), it is inappropriate to use the test for diagnostic purposes: that is, to determine an examinee's level of performance on a single "pure" variable. This is because of the fact that two different examinees may obtain identical scores by marking correctly the same number of different items. This possibility obviously is greatest for scores midway between the maximum and the minimum. With number-right scoring, it is nonexistent for zero scores or perfect scores or for single-item tests. It should perhaps be noted that homogeneous items can be at very different levels of difficulty in any given population of examinees. An illustration of this point was given by Davis (1946, p. 25).

---

[1] A ***representative sample*** consists of a relatively small number of units (measures, objects, tasks, etc.) drawn from all units defined as being within precise boundaries (called a population or universe) in such a way that the sample duplicates a set of specified characteristics of the population.

From an indefinitely large population of tasks that lie within specified boundaries, successive samples may be drawn at random in such a way that each task has a known probability of being selected for inclusion in any given sample. Each sample of tasks has the same probability of being selected. There is no need to put the tasks drawn in any given sample back into the population before another sample is drawn at random because the population of tasks is conceived of as being inexhaustible; that is, indefinitely large. We refer to a sample of this kind as a ***simple random sample without replacement***. A sample that has characteristics that represent the average of the characteristics of an indefinitely large number of samples of this kind is called a ***representative sample*** of the population of tasks. Any particular simple random sample may differ by chance from the characteristics of its parent population of tasks, but a representative sample cannot do so. It is evident, then, that the concept of a representative sample is theoretical. But this fact does not diminish its usefulness for purposes of discussion; indeed, the fact may enhance its usefulness for such purposes.

Since a population of tasks is often an aggregation of tasks of divergent characteristics that lie within the boundaries specified for the population (often called a ***domain*** by test constructors), we may reduce the likelihood that tasks of certain characteristics will, by mere chance, be greatly overrepresented or underrepresented in a sample by using the method known as ***stratified random sampling without replacement***. This is accomplished by identifying tasks in the population that exhibit differing characteristics and by grouping these into more-or-less homogeneous strata or subpopulations. The same predesignated proportion of the tasks in each subpopulation is then drawn by a simple random sampling procedure and the resulting subsamples are merged together to form a stratified random sample without replacement. A sample that has characteristics that represent the average of the characteristics of an indefinitely large number of samples of this kind is also a ***representative sample*** of the population of tasks. If each subpopulation is rather homogeneous with respect to the characteristics of the tasks within it, and if the characteristics of the subpopulations differ rather markedly from one to another, the sampling error of stratified random samples is likely to be smaller than the sampling error of simple random samples. That is the principal reason for obtaining stratified random samples instead of simple random samples.

The implication for the preparation of homogeneous items for a multi-item diagnostic test is that each item must measure only one "pure" variable, plus error, or the same weighted combination of two or more "pure" variables, plus error. In either of these cases, the item scores would be found to measure, at a preselected level of significance, the same dimension except for errors of measurement and for differences of origin and of units of measurement (Lord, 1973b; Villegas, 1964). Kriewall (1972) expressed this idea in discussing item selection for criterion-referenced tests when he wrote:

> The item-sampling model described here as the paradigm for CRT construction is one of the simplest models. It places no conditions on the items except, to preserve score meaning, all items must share at minimum the objective attributes which serve to characterize an L(earning) O(bjective) (p. 16).

Although this model imposes no uniformity on item format, level of difficulty, or item-generating procedures, it is basically a very restrictive definition of a criterion-referenced test. It means that all criterion-referenced tests would have to be comprised of homogeneous items and would, therefore, be useful for diagnosis to estimate the level of performance of each individual on the single variable measured by scores on each test.

As noted above, Nitko's (1970) definition of a criterion-referenced test would permit it to be comprised of heterogeneous items. As the degree of heterogeneity increases, total scores derived from such a test become more and more inappropriate for diagnostic purposes. But they may be entirely appropriate for survey purposes to estimate the proportion of a population of items constituting a domain, or area of subject-matter content, that can be answered correctly by an examinee. Glaser and Nitko (1971) defined a criterion-referenced test as a

> . . . measuring instrument deliberately constructed to yield measurements that are directly interpretable as performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain and such measurements are directly referenced to this domain for each individual measured (p. 653).

Nitko (1970) gave an example of a criterion-referenced test that covered a rather widespread domain; namely, elementary-school geometry. He wrote:

> . . . a score of 30 might mean that, along with a number of lower-level behaviors, the student is able to identify pictures of open continuous curves, lines, line segments, and rays; can state how these are related to each other; and can write symbolic names for specific illustrations of them. He can identify pictures of intersecting and nonintersecting lines and can name the point of intersection. This score would also mean that the student could *not* demonstrate high-level behaviors, such as identifying pictures that show angles; naming angles with three points; identifying the vertex of a triangle and an angle; identifying perpendicular lines; using a compass for bisection or drawing perpendiculars; and so on (p. 38).

In the absence of statistical evidence that all of the items in this test of elementary geometry measure the same dimension, plus error, it seems reasonable to conclude that they are somewhat heterogeneous in content and that each of them is referenced to a task that is part of the domain covered. If we assume that they constitute a stratified random sample, without replacement, of an indefinitely large population of tasks that constitutes this domain, it would be legitimate to use the ratio of the number of items marked correctly to the number of items in the test as a crude estimate of the proportion of items in the population that the examinee would mark correctly (*not* the proportion of tasks in the population that he could perform satisfactorily). If the probability of guessing correctly in marking responses to the items were zero, it would be proper to use the ratio specified as a crude estimate of the proportion of tasks that the examinee could perform satisfactorily. It is this type of interpretation of the scores from criterion-referenced tests that Harris and Stewart (1971) had in mind when they wrote: "We conceptualize a pure criterion-referenced test as one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of the performances in that population at which the student can succeed." This definition of a criterion-referenced test, like those of Nitko (1970, p. 38) and of Glaser and Nitko (1971, p. 653), permits construction of either a diagnostic instrument (made up of homogeneous items) or a survey instrument (made up of heterogeneous items). However, unlike the definition given by Glaser and Nitko (1971, p. 653), it fails to specify that the sample of production tasks be a representative sample of the population of production tasks that defines the domain being measured. In practice, stratified random sampling without replacement is ordinarily preferred in test construction in order to avoid traces of memory variance or variance attributable to differential tendencies among examinees to look up the answers to items after they have taken a test. These considerations assume greater importance as the number of items in a test is decreased; and criterion-referenced tests tend to be short in many of their applications.

Two additional points should now be mentioned with respect to the legitimate interpretation of a score of 30 on the test described by Nitko (1970, p. 38) provided by forming a ratio with 30 as the numerator and the (unspecified) number of items in the test as the denominator for use in estimating the proportion of the parent population of *items* that would be marked correctly by an examinee. First, if the reliability coefficient of the test scores can be obtained in the group of examinees for whom interpretations of scores are to be made, the obtained score in the numerator may be replaced with an estimated true score to secure a better estimate of the proportion of the parent population of items that the examinee would mark correctly.

Second, if the probability of guessing the correct answers to *items* is zero, this proportion would represent a good estimate of the proportion of *tasks* in the parent population of tasks that the examinee would be able to perform correctly. In practice, the probability of guessing correct answers to items is not likely to be zero, especially if the items constitute a pretest or if the amount or quality of instruction provided prior to a posttest is inadequate. With multiple-choice items, the probability of marking any one item correctly by guessing equals $1/c$, where c is the number of choices in the item. It seems likely that examinees rarely guess among all choices in an item. Instead, they are able to eliminate one or more choices as incorrect, and they may guess among the remaining choices. The probability of marking an item correctly by guessing under these circumstances is $1/(c-x)$, where x is the number of choices that can correctly be identified as incorrect. Sometimes examinees mark a response on the basis of nonchance factors that are, in well-constructed tests, unrelated to the correctness or incorrectness of a choice. For example, some examinees will systematically mark the longest choice or the most precise-sounding choice among those that they cannot eliminate as definitely wrong. In these circumstances, other examinees mark the choice in the position (i.e., A, B, C, D, etc.) that they marked as correct least recently. Since experienced item writers or editors deliberately introduce some long and precise-sounding distracters into their items and careful test editors always arrange to have all choice positions appear about an equal number of times per test and in random order, use of these nonchance factors is no more effective than sheer guesswork in raising an examinee's score on a well-constructed test.

With free-response items, it is often thought that the probability of guessing correct answers is so close to zero that it may be ignored, but this is often not so. Consider, for example, an item drawn from a criterion-referenced spelling posttest. A short sentence containing the word "receive" is dictated and the examinees are asked to write the word "receive" on line 12. Most of the examinees know that the word is spelled "receive"; some guess between two alternative spellings, "receive" and "recieve"; two pupils who missed the instruction have to guess among other phonetic possibilities like "receeve," "reseeve," etc. In summary, examinees do not usually construct free responses at random or from a very large number of possibilities. In practice, answering free-response items is often a process of knowing the correct response and supplying it or of selecting a response from two or more that are brought to mind by the stimulus.

From the foregoing discussion, it is clear that no wholly adequate practical procedure is available, or can be made available, for estimating the proportion of *tasks* in a parent population of tasks that can be performed by an examinee; we can, strictly speaking, infer only what proportion of *items* in a parent population of items he can answer correctly unless the tasks themselves can be used as items. However, the conventional correction for chance success will in most cases be found to yield scores for examinees that are closer to guessing-free scores than are number-right scores. Davis (1959), Little and Creaser (1966), and Cross (1973) have provided data relevant to this point, and Lord (1973a) has argued that, when appropriate directions for administering tests are used, formula scores (used to correct for chance success) are better estimators of the rank order of examinees in the trait measured by a test than are number-right scores if examinees omit items when the test is administered with directions appropriate to formula scoring.

PREPARING ITEMS FOR CRITERION-REFERENCED TESTS

A prerequisite for the preparation of items for any well-constructed instrument for evaluating achievement has always been a carefully prepared and highly detailed outline of the population of knowledge (facts and understandings), skills, or feelings to be assessed. Within this outline, each topic, cell, or block must be weighted to make it contribute a thoughtfully determined proportion of the variance of the total score yielded by the test in a representative sample of examinees drawn from the population to be assessed with the instrument. This makes the total scores appropriate for use as survey measures of a multifaceted domain. Nitko's example of the test of elementary geometry (Nitko, 1970, p. 38) provides an illustration of a test for which total scores may be used in this way.

If subscores from a survey test are to be used for diagnostic purposes, the items in each part should constitute a representative sample of a homogeneous stratum, cell, or block in the population of objectives to be measured. As already noted, unambiguous interpretation of scores for diagnostic purposes requires that such scores must be based on homogeneous items.

Procedures that are appropriate for generating items capable of eliciting examinee behaviors that literally constitute overt manifestations of the feelings, skills, and knowledge (facts and understandings) that make up the population of objectives necessarily vary with the nature and level of the content to be measured. But certain fundamental principles are widely observed. For multiple-choice items, the following may be mentioned:

1. The keyed response must be an adequate correct response—not merely the best of the responses included. In the following item, the keyed response, "empty," is inadequate even though examinees who recognize its inadequacy will mark it as correct.

   To pour means to
   A drive.
   *B empty.[2]
   C lack.
   D hurt.

2. All distractors must be clearly incorrect or (in best-answer items) generally accepted by informed authorities in the field as less adequate answers than the keyed response. The following reading-comprehension item was based on a short passage that described the appearance and functions of a post office. To measure comprehension of the passage, certain words were deleted and the examinees were asked to select the best of two words to fill in the blank space. The words deleted and the distractors used in each item were determined by a set of objective rules for selecting words at random from the passage so that the subjective judgment of the item writer would not be involved. One item read:

   Many people __________ here.
   A post
   *B work

The rules for selecting distractors for these items unfortunately did not include the important constraint that no distractor could be a defensible answer; and in the case of this illustrative item the only distractor is so defensible. In general, sensible restraints should be placed on the generation of distractors, as shown by the third general principle.

3. Distractors should be as attractive as the psychological context of the item permits and should be as nearly equally attractive to examinees in the target population as possible; that is, each distractor should attract as nearly as possible the same proportion of those examinees who cannot identify the correct answer (Horst, 1933).

---

[2] In the above item, and in subsequent illustrative items, the keyed choice is identified by an asterisk.

When a set of 3-choice word-matching exercises was being constructed for use in a diagnostic reading test for elementary-school pupils, the word "kitchen" was one chosen from the group of seven-letter words commonly used in basic readers. Two distractors were needed. The rules for obtaining distractors for this item specified that distractors must (a) be made up of words or letter combinations that were seven letters in length but must not be the stimulus word; (b) begin and end with the same two letters as the stimulus word; and (c) have the same upper and lower profiles as the stimulus word. A skilled item writer exercised her best professional judgment in producing distractors that resulted in the following item, which requires that the examinee identify the lettered word that is the same as the word at the left.

kitchen	kitchen *A	kitcken B	kitehen C

When this item was administered to 140 pupils in grades 5 and 6, the data (presented in Table 1) showed that the item was easy (137 examinees or 97.9 per cent, marked it correctly) and the biserial correlation coefficient between scores on it and scores on a group of 99 mechanics-of-reading items in which the item was *not* included was .37. The choice-by-choice data indicate that two pupils marked it incorrectly; both of these were in the lowest fifth of the score distribution for the 99-item test. One pupil

Table 1. Item-Analysis for A Word-Matching Item (N = 140)*

	Number of Examinees					
Fifth of Total-Score Distribution	Choice					
	A#	B	C	Omitted	Not Read	Total
Lowest	26	1	1	0	0	28
2nd	27	0	0	1	0	28
3rd	28	0	0	0	0	28
4th	28	0	0	0	0	28
Highest	28	0	0	0	0	28
Total	137	1	1	1	0	140

$P_A = .979$ $r_{bis} = .372$

*These data are provided by a proprietary item-analysis program largely written by Daniel Ashler. The program is not available, but it may be used by arrangement with the Center for Research in Evaluation and Measurement (CREM), Graduate School of Education—C1, University of Pennsylvania, Philadelphia, Pennsylvania 19174. The item for which the data are shown was not included in the criterion variable used in the item analysis.

# Keyed response

in the second-lowest fifth of this distribution read but failed to mark an answer to the item. No pupil failed to read the item, though it was 93rd in a set of 100 items administered to them.

This illustrative item is given largely to show the kinds of specifications that should be provided for item writers. Except for items measuring the most basic skills (like addition of single-digit numbers), item construction involving only the blind following of rules is nugatory. Effective distractors must not involve tricky or deceptive concepts, but they should include natural misconceptions. For example, an item for college students intended to measure knowledge of the meaning of "pedantic" should include a distractor like "having feet" or "footed" because superficial (and, in this case, irrelevant) knowledge of the fact that "pedes" is the Latin word for "feet" often causes "footed" to be chosen as the synonym for "pedantic" by examinees who do not know the meaning of the word but hope to give the impression that they do by using extraneous information and verbal-reasoning skill.

After appropriate sampling procedures have been used to identify the tasks in the parent population that are to be tested and after objective and consistent rules have been followed that describe the specific characteristics of the item stems, keyed responses, and distractors, the validity

Table 2. Paths to Answers Expected of Typical High-School Juniors and Seniors to a Five-Choice Arithmetic-Reasoning Item

35. A radio costs \$60. A man pays \$9 down and the remainder in 9 monthly payments of \$7 each. This method of payment increases the cost of the radio by what percent?

Choice Letter	Choice	Mental Process Expected
A	5%	$\frac{7 \times 9 - 60}{60} = .05 \times 100 = 5$
B	6%	$\frac{7 \times 9 + 9}{7 \times 9 + 9 - 60} = 6$
C	8⅓%	$\frac{60}{7 \times 9 + 9} = .8\tfrac{1}{3} \times 10 = 8\tfrac{1}{3}$
D	16⅔%	$\frac{7 \times 9 + 9 - 60}{7 \times 9 + 9} = .16\tfrac{2}{3} \times 100 = 16\tfrac{2}{3}$
*E	20%	$\frac{7 \times 9 + 9 - 60}{60} = .20 \times 100 = 20$

*Keyed response

of each item for measuring the precise criterion performance to which it is referenced often depends mainly on the psychological insight and ingenuity of the item writer except, as noted, in the preparation of items that test simple skills and associations.

Table 2 shows an arithmetic-reasoning item for which the expected path to each of the five responses is given in compact form. Operating within the framework of rules governing the sampling from the population of objectives and the construction of distractors, the item writer hypothesized four incorrect ways of solving the problem likely to be used by examinees (in the population to be tested) at each of four levels of arithmetic-reasoning ability. The four distractors resulted. Sometimes items of this kind have been given as free-response items and popular incorrect answers obtained by examinees at four levels of ability have been used as distractors.

Gross grammatical faults occasionally cause distractors to be unattractive. For example, in the following item, choice D is rarely chosen by examinees because they notice at once that it does not grammatically fit the stem.

A closed plane figure with six sides is called a
A cube.
B pentagon.
*C hexagon.
D octagon.

Any characteristic of an item that leads examinees to avoid a distractor or to choose the correct answer without using the knowledge or skill that it is intended to measure is called a specific determiner. In the following item, examinees who do not understand the point that the item was written to test are drawn to the correct answer (choice B) because of the prominence of the word "better" in the choice.

The most important advantage man has over other animals is
A the habit of walking on two legs.
*B better ways of thinking.
C automobiles and many conveniences.
D a keen sense of sight and hearing.

In the context, the word "better" is a specific determiner. Examinees who do not understand what the item is testing will tend to avoid distractor C because "automobiles and many conveniences" are plurals following the singular verb "is." Such lack of agreement is also a specific determiner. The effects of these specific determiners are complicated by the fact that the correct response includes a plural predicate nominative, "ways," but this may be accepted as a collective noun more readily than "automobiles" or "conveniences" in distractor C. It should be noted that examinees need not understand formal grammar to be influenced by any disagreement between the subject and verb in the stem of the item and the predicate nominatives that follow in the choices. They need only be sensitive to the "feel" or "sound" of the sentences involved.

Many other kinds of specific determiners appear in items. The presence of "always," "never," and similar absolutes in a choice immediately signals the examinee that the choice is not likely to be the correct answer. The presence of absolutes like this in the stem of a true-false item is widely recognized by the examinees as an indication that the item is probably keyed as "false."

4. Choices for an item should be logically coordinate and distractors should not overlap each other or be related in a way that allows one or more distractors to be eliminated by an examinee who is test-wise and can reason well but has no information or skill in the variable that the item is intended to measure. In the following item from a Sports and Hobbies Test, choice A was keyed as correct.

Quail are usually found
*A near grain fields.
B in swamps.
C running across the road.
D in marshes.

Choices B and D are so nearly alike in meaning that an alert examinee without knowledge of quail or their habitats recognizes that if B were keyed as correct, D would also have to be, or vice versa. Since he knows that only one keyed response is provided in each item, he realizes that neither B nor D can be the correct answer. Choice C differs in level of generality ("the" road) from the others. Hence, he recognizes that it is not logically coordinate with the other choices and regards it as an unlikely correct answer. Without using any information about quail, the alert examinee has correctly identified choice A as the keyed response.

This item was written to measure information about a perfectly legitimate objective in a test about sports and hobbies. The stem of the item, the keyed response, and distractor B cannot be faulted, but the item was rendered invalid by the inclusion of distractors C and D.

Another example of how logical interrelationships of distractors can render an item invalid when examinees are told to mark only one choice as correct or as the "best answer" is the following:

In 1960, the population of Holland was over
A 2 million.
B 4 million.

C 8 million.
*D 10 million.
E 20 million.

Choice D was keyed as correct. But it is apparent that choices A, B, and C are equally correct. To maintain that choice D is the "best answer" is not defensible. With a different wording of the stem, this item of information about Holland could be tested adequately by using the same choices. Variants of this fault in item writing are commonly found.

A great many rules for item writing have been suggested, but most of them can be subsumed under the four that have been stated and illustrated in the foregoing pages. For example, Masonis (1971) listed 47 principles for writing multiple-choice items. Studies of violations of some of these principles have been published (Board & Whitney, 1972; Chase, 1964; Dunn & Goldstein, 1959; McMorris, Brown, Snyder, & Pruzek, 1972; Terranova, 1969; Williamson & Hopkins, 1967). Results of some of these studies have been summarized by Pyrczak (1972). An interesting experiment on the effect of four sets of instructions to item writers on the difficulty level and homogeneity of items was reported by Baker (1971). She provided four groups of inexperienced item writers drawn from the same population with (a) a general objective for writing items; (b) a behaviorally stated objective for writing items; (c) a behaviorally stated objective for writing items plus a sample item; and (d) a behaviorally stated objective for writing items plus five statements designed to specify the item form desired. When items written by the four groups were compared, it was found that instruction (a) produced the most difficult items and instruction (c) produced the easiest items. The comparison also showed little difference in the average intercorrelations of the items produced under the four different instructions. These results are not in disagreement with our expectation that differences in the psychological insight, conscientiousness, and experience of the item writers have greater effect on the quality and other characteristics of test items than do other variables. Much more experimental research is needed in this area.

SELECTING ITEMS FOR CRITERION-REFERENCED TESTS

When a pool of items is available from which a designated number is to be selected for use in a given test form, test editors are guided in their selection by four basic principles that have been widely used for 25–30 years:

1. The items in an achievement test should constitute as nearly as possible a representative sample of the population of items that define the domain to be measured. As indicated previously in this paper, the best way to accomplish this would be to draw a stratified random sample, without replacement, from the population of items. Needless

to say, the latter must measure precisely the behaviors, or task performances, that make up the criterion variable that is to be measured; and these task performances must be reflected in the population of items in their proper proportions to one another. Adherence to principle 1 will, under these conditions, maximize the content validity of achievement tests and is of overriding importance in the construction of tests of this type, which includes criterion-referenced tests, whether they are to be used for diagnostic or survey purposes and whether they are to be used immediately following a single unit of instruction or following a lengthy course of study.

2. The items in a predictor test, such as an aptitude or selection test used in education or industry, should constitute the set (drawn from the population of items that define the domain to be tested) which best predicts scores on the designated criterion variable in samples of examinees like those to whom the test will be administered. This prediction capability must be demonstrated at critical score levels, which are immediately adjacent to a cutting point if one is used.

The pool of items prepared for use in constructing an aptitude or selection test is often deliberately made to measure a representative sample of the population of tasks that comprises performance in a course of study or a specific job. This is because simulations of such tasks, or work samples, often turn out to be the best predictors and because court decisions may require aptitude or selection tests to demonstrate content validity as well as criterion-related validity if their use is to be adjudged legal. However, the items selected for use in a predictor or selection test need not constitute a representative sample of the items that define the domain to be tested. Since this paper deals with the preparation of criterion-referenced tests, principle 2 will not be discussed further.

3. The items in an achievement test should, within the constraint imposed by principle 1, make up as efficient a measuring instrument as it is possible to produce.

There are many criteria by which the efficiency of a test may be judged. Two of those commonly used in the past are (a) the extent to which raw scores yielded by a test differentiate among all the examinees tested; and (b) the extent to which raw scores yielded by the test differentiate examinees in one category of test scores from those in a second category (without considering differentiations among the examinees within either category).

It is easy to show that maximum differentiation among all examinees in a group is obtained when the distribution of raw scores is rectangular across the entire range of possible scores. A K-item test can provide K + 1 score categories from a score of 0 through a score of K. In practice it is not possible to count on obtaining completely rectangular distributions of raw scores, but they can be approached as closely as the intercorrelations of the individual items will permit if each item in the test is of a difficulty level such that one-half of the examinees in the population mark it correctly. This holds true when the items are scored in such a way that one point is given for each correct answer marked and 0 is given for each incorrect answer marked or for each omission and when the average of the item intercorrelations (phi coefficients) falls in the range from .00 to .33. The latter value is rarely if ever exceeded because single items ordinarily yield highly unreliable scores. However, data reported by Scandura and Durnin (1971) show higher reliability coefficients for a few single items testing specific behaviors (the use of rules in solving easy arithmetic problems) that had been taught and practiced just prior to the testing. Further research is needed on the intercorrelations of very homogeneous items administered to examinees for whom they are very easy. As the average item intercorrelations rise above .33, closer and closer approximations to rectangularity of raw-score distributions can be obtained by increasing the spread of item difficulties around the 50-percent difficulty level.

In situations where a group of examinees is to be divided into two categories—examinees whose raw scores on a test are at or above a designated cutting point (or "passing mark," as it is sometimes called) and examinees whose raw scores are below the cutting point—it has been shown that maximum differentiation between examinees in one category and those in the other category will be obtained if the distribution of raw scores is such that half of the examinees fall into each of the two categories. This situation may be brought about if each item in the test is marked correctly by half of the members of the population (of which a representative sample has been tested) whose true competence level is represented by the cutting point. That is to say, each item is of 50 percent difficulty for those examinees in the true-score distribution whose test scores correspond to the cutting point.

It is obvious that an index of efficiency for criterion-referenced tests that would indicate the extent to which a given test differentiates examinees whose scores are at or above the cutting point (those who have "mastered" the content measured) from examinees whose scores are below the cutting point (those who have not "mastered" the content measured) would be of interest and value to test users and research workers. Harris (1972) has shown that the squared product-moment biserial (point-biserial) correlation coefficient between test scores (based on free-response items scored 1 for a correct response and 0 for an incorrect response or an omission) and a variable created to represent membership in the two categories (say, 1 for examinees who have scores at or above the cutting point and 0 for examinees who have scores below that point) constitutes such an index. He notes that this index takes its maximum value in symmetrical distributions when about half of the examinees fall into each of the two categories. It becomes larger as the distribution of test scores approaches rectangularity.

The practical question that now arises is "Can a test constructor markedly increase the efficiency of a criterion-referenced test, as measured by the Harris index of test efficiency, by selecting items in conformance with principle 1 that are at, or close to, the 50 percent difficulty level for those examinees in the true-score distribution whose scores correspond to the cutting point?" Unless the test constructor has a pool of items that includes a large number of equivalent items testing each of the specific performances that his test is to measure, it is unlikely that enough items at or close to the required difficulty level will be available. Thus, because of the overriding importance of principle 1, it is doubtful that increases in test efficiency of practical significance will usually be obtained by selecting items according to difficulty for criterion-referenced tests. It is theoretically possible, and research to ascertain the value of the procedure under practical circumstances is needed.

4. Choice-by-choice item-analysis data should be used as a basis for editing and revising items for achievement, aptitude, and selection tests.

Tabulations of the percentage of examinees who marked each choice in objectively scorable items should be obtained separately for examinees in a high-capability and a low-capability subsample of examinees in a representative sample of the population in which the test is to be used. In addition, the percents of examinees who omitted the item or who failed to reach the item in the time limit should be tabulated in each of the subsamples. These data are likely to prove useful in detecting items that are clearly defective. Keyed choices that are not marked as correct by a larger percent of the high-capability examinees than by the low-capability examinees and distractors that are not marked as correct by a larger percent of the low-capability examinees than by the high-capability examinees point to the need for revising or discarding items. Illustrations of the use of choice-by-choice data in editing test items and of its effects have been provided by Davis (1951, pp. 305–308).

It should be noted that the criterion variable used for identifying high-capability and low-capability examinees may be the total score on the test comprised of the items being studied or the total score on a parallel form of the test administered to the same examinees with a separate time limit. For purposes of item analysis, time limits should be so liberal that all, or almost all, of the examinees have time to read every item. When parallel forms A and B of a tryout test are used, the total scores on form A should be used as the criterion variable to establish subsamples for studying the items in form B. The converse of this procedure is used for studying the items in form A. This procedure avoids spurious inflation of differences between the percents of examinees who mark each choice in the high-capability and low-capability subsamples; it also eliminates the troublesome problem of differential inflation of item-score versus total-score correlation coefficients. The procedure is particularly valuable if the number of items in a test is small. It takes the same amount of testing time and it costs little more than conventional procedures if optical-scanning and computer facilities are available. If two parallel forms are not used, and scores from a given item are correlated with the total scores in which they are included, the resulting correlation will be spuriously high and data to make the proper correction statistically (Davis, 1958, Equation 4) are almost impossible to obtain.

The variable used to establish the low-capability and high-capability subsamples can also be a set of external criterion scores. Their use is especially desirable in selecting items for predictor tests of any kind. In fact, for tests of these kinds, item-analysis data based on both total-score and external-criterion variables should be obtained. One of several procedures for maximizing test validity by using both item-score versus total-score and item-score versus external-criterion-score correlation coefficients may then be used (Davis, 1947, pp. 20–23; Gullikson, 1950; Horst, 1936).

In the construction of achievement tests, item-score versus total-score correlation coefficients of various types, t tests, and other statistics often used as "item-discrimination" indices should be employed cautiously if, indeed, they are employed at all. Davis (1946, pp. 19–20) stressed this point as early as 1946 and later wrote (Davis, 1952):

> For achievement tests, great care must be exercised that items judged unacceptable by subject-matter experts be excluded and that the final form preserve the balance among topics specified in the test outline. Then, too, proper regard for the shape of the distribution of item difficulties must be observed, as noted earlier in this article. The value of item-discrimination indices must always be considered in the light of the adequacy of the criterion variable, the purpose for which the test is to be used, and the way it serves that purpose . . . the usefulness of item-discrimination indices is often smaller than is commonly supposed (pp. 116–118).

Unfortunately, item-analysis data have frequently been used mechanically to select items for the final form of an achievement test mainly on the basis of item-score versus total-score discrimination indices of one sort or another. In the construction of criterion-referenced tests of the survey type, their use is highly inadvisable; for building homogeneous tests of a single performance objective for diagnostic purposes, their use may be helpful in making the items in the final form of the test more homogeneous with respect to the function that it is desired to measure; that is, the set of items having the highest item-score versus total-score discrimination indices may constitute more nearly a "pure" test than the tryout test from which the set was chosen.

The use of posttest-pretest difference scores for a sample of examinees as criterion scores for item analysis has been described by Cox and Vargas (1966) and by Hambleton and Gorth (1970). Wholly aside from the fact that such difference scores are notoriously unreliable because of the usual high correlation between pretest and posttest scores, the selection of items that show the highest item-score versus difference-score correlations must be done within the overriding constraint imposed by principle 1; otherwise, the content validity of the final form of a test may be seriously impaired. For examinees who have been given the same instruction between pretest and posttest administrations, the major variable measured by posttest-pretest difference scores is capacity to learn the material taught.

If items are selected for the final form of a test on the basis of high correlations with this variable, variation among examinees with respect to the capability they have developed to learn the kind of material taught becomes the "criterion" to which the items are referenced. It is precisely to avoid this outcome that caused Glaser and Nitko (1971) to define a criterion-referenced test as a "measuring instrument deliberately constructed to yield measurements that are directly interpretable as performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual (p. 653)." In other words, each examinee is to be measured to discover the extent to which he has attained the objectives of instruction and not to discover how well he compares with other examinees on capacity to learn the material taught.

CHOICE OF A CUTTING SCORE

A cutting point, or "passing mark," may be established for an achievement test by any test user in terms of standards that he deems useful and meaningful to him and to the examinees who have taken or will take the test. The setting of a passing mark for a criterion-referenced test is, therefore, both optional and highly subjective. If, however, a cutting point is intended to mark the dividing line between examinees who have "mastered" the content of the population represented by the sample of items that make up the test and examinees who have "not mastered" that content, certain logical considerations must be taken into account. Strictly speaking, mastery is defined as complete knowledge, skill, or control; so "partial mastery" is as self-contradictory a phrase as "partial uniqueness." The term "mastery," therefore, should be used to describe the status of only those examinees who, it may be inferred, can mark correctly all items in the population of which the subset that makes up a criterion-referenced test is a representative sample. Theoretically, this requires that only examinees who obtain a perfect score on the test can be regarded as having mastered the content of the population of items. The reason for this is that, mathematically, we regard it as impossible for an examinee who has complete knowledge of all items in the population to mark incorrectly any item drawn from that population.

Given this result, we may prefer to set a cutting point at some level of competence in the domain measured that is lower than mastery but is high enough to meet practical requirements. For example, an elementary-school pupil learning the fundamental operations of arithmetic should probably be required to attain a level of competence close to mastery (say, perhaps, a level represented by the ability to mark correctly at least 95 percent of the items in the population represented by the test) because many skills to be learned depend on others that have been learned previously. In social studies, however, a pupil can learn about communities of various types without having to depend heavily on information about families that he has been expected to learn previously. Therefore, the level of competence required in some subject-matter fields may, as a practical matter, be established farther below mastery level than in others.

To take this factor into account, Table 3 has been constructed to show estimates, based on the number of items in a test that an examinee has marked correctly, of the probability that an examinee's true level of competence is at or above the levels represented by complete knowledge of .99, .95, .90, .85, or .80 of the items in the population represented by the test. This has been done for 5-item tests, 12-item tests, and 20-item tests. Scores from most criterion-referenced tests fall within those limits, but the method used to construct Table 3 is general and can be used to extend the table to care for any test length. The values shown in Table 3 were obtained by application of Bayes' theorem. In doing this, it was assumed that the test user would have no data to employ in estimating an examinee's true competence level other than his test score (obtained by giving one point for each item answered correctly and zero points for each item answered incorrectly or omitted).

Consider the data in Table 3 pertaining to a 12-item test. The probability that an examinee who obtained a score of 12 has a competence level at or above .99 is only .2300, but this is much higher than the probability that an examinee who obtained any lower score has a true competence level at or above .99. It will be noted that the category for which .99 is the lower limit is virtually the same, for all practical purposes, as mastery. As the minimum level of competence in the five categories is lowered to .80, the probability that an examinee who obtained a score of 12 has a true competence level at or above .80 becomes .9524. In fact, the probability that an examinee who obtained a score of 10 has a true competence level at or above .80 is as high as .5173.

We can specify cutting scores for the 12-item test if we designate the lowest acceptable competence level that we are willing to accept and the proportion of correct classifications of examinees (below, at, or above

Table 3. Probabilities and Cutting Scores for Selected Levels of Competence, Obtained Scores, and Test Lengths

Competency Level at or Above	Obtained Score					Cutting Score When Probability of Correct Categorization of Examinee Is .5000	.8500
	5-Item Test						
	5	4	3	2			
.99	.1135	.0028	.0000	.0000		5	5
.95	.3085	.0385	.0026	.0001		5	5
.90	.5006	.1235	.0174	.0016		5	5
.85	.6459	.2343	.0505	.0067		5	5
.80	.7539	.3558	.0940	.0186		5	5
	12-Item Test						
	12	11	10	9			
.99	.2300	.0140	.0008	.0000		12	12
.95	.5509	.1568	.0312	.0043		12	12
.90	.7784	.4021	.1493	.0398		12	12
.85	.8950	.6204	.3273	.1292		11	12
.80	.9524	.7790	.5173	.2660		10	12
	20-Item Test						
	20	19	18	17	16		
.99	.3441	.0348	.0033	.0002	.0000	20	20
.95	.7269	.3202	.1054	.0255	.0048	20	20
.90	.9135	.6618	.3818	.1720	.0619	19	20
.85	.9749	.8587	.6537	.4135	.2165	18	20
.80	.9939	.9479	.8359	.6501	.4364	17	19

this level) that is desired. If the lowest acceptable competence level is .80 and the desired proportion of correct classifications is .8500, the lowest acceptable cutting score is 12. On the other hand, if we are willing to accept a proportion of correct classifications as low as .5000, the cutting score can be set at 10.

The sections of Table 3 pertaining to 5-item and to 20-item tests can be interpreted in the same manner. By and large, the data show that if a test user wants to identify examinees whose competence levels are .90 or higher and wants to make no fewer than 85 correct classifications out of every 100 that he makes, he must set the cutting score for a test of up to 20 items at the maximum possible score. On the other hand, if he requires less accuracy in his classifications and is satisfied with a cutting score that represents a level of competence considerably below the mastery level for the domain being measured, lower cutting scores can be used.

It should be noted that the mathematical model that underlies the data in Table 3 does not allow for carelessness, clerical error, etc., on the part of examinees. It should also be noted that the probabilities are different from what they would be if additional information about an examinee's competence other than his test score were employed by the test user. If highly relevant information were available and were properly used, true competence level for an individual examinee could be estimated with higher probability than is shown in Table 3. Consequently, cutting scores different from those shown could be employed.

REFERENCES

Baker, E.L. The effects of manipulated item writing constraints on the homogeneity of test items. *Journal of Educational Measurement*, 1971, 8, 305–309.

Board, C., & Whitney, D.R. The effect of selected poor item-writing practices on test difficulty, reliability, and validity. *Journal of Educational Measurement*, 1972, 9, 225–233.

Chadwick, E. Statistics of educational results. *The Museum, Quarterly Magazine of Education, Literature, and Science*, 1864, *3*, 480–484.

Chase, C.I. Relative length of option and response set in multiple-choice items. *Educational and Psychological Measurement*, 1964, *24*, 861–866.

Cox, R.C., & Vargas, J.C. A comparison of item-selection techniques for norm-referenced and criterion-referenced tests. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

Cross, L.H. An investigation of a scoring procedure designed to eliminate score variance due to guessing in multiple-choice tests. Unpublished doctoral dissertation, University of Pennsylvania, 1973.

Davis, F.B. *Item-analysis data: Their computation, interpretation, and use in test construction*. Cambridge, Mass.: Harvard Graduate School of Education, 1946.

Davis, F.B. *The AAF Qualifying Examination*, Washington: Government Printing Office, 1947.

Davis, F.B. Item selection techniques. In E.F. Lindquist (Ed.), *Educational measurement*. Washington: American Council on Education, 1951.

Davis, F.B. Item analysis in relation to educational and psychological testing. *Psychological Bulletin*, 1952, 2, 99–212.

Davis, F.B. A note on part-whole correlation. *Journal of Educational Psychology*, 1958, *49*, 77–79.

Davis, F.B. Use of correction for chance success in test scoring. *Journal of Educational Research*, 1959, *52*, 279–280.

Dunn, T.F., & Goldstein, L.G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. *Educational and Psychological Measurement*, 1959, *19*, 171–179.

Gagné, R.M. *Psychological principles in systems development*. New York: Holt, Rinehart, & Winston, 1962.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, *18*, 519–521.

Glaser, R., & Klaus, D.J. Proficiency measurement: Assessing human performance. In R.M. Gagné (Ed.), *Psychological principles in systems development*. New York: Holt, Rinehart, & Winston, 1962.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational measurement*. Washington: American Council on Education, 1971.

Gulliksen, H.O. *Theory of mental tests*. New York: Wiley, 1950.

Hambleton, R.K., & Gorth, W.P. Criterion-referenced testing: Issues and applications. A paper presented at the annual meeting of the Northeastern Educational Research Association, Liberty, N.Y.: 1970. (ED 060 025, MF and HC available from EDRS)

Harris, C.W. An index of efficiency for fixed-length mastery tests. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Harris, M.L., & Stewart, D.M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Horst, A.P. The difficulty of a multiple-choice test item. *Journal of Educational Psychology*, 1933, *24*, 229–232.

Horst, A.P. Item selection by means of a maximizing function. *Psychometrika*, 1936, *1*, 229–244.

Kriewall, T.E. Aspects and applications of criterion-referenced tests. Downers Grove, Ill.: Institute for Educational Research, 1972. (ED 063 333, MF and HC available from EDRS)

Lindquist, E.F. (Ed.) *Educational measurement*. Washington: American Council on Education, 1951.

Little, E., & Creaser, J. Uncertain responses on multiple-choice examinations. *Psychological Reports*, 1966, *18*, 801–802.

Lord, F.M. Formula scoring and number-right scoring. Unpublished paper, 1973. (a)

Lord, F.M. Testing if two measuring procedures measure the same dimension. *Psychological Bulletin*, 1973, *79*, 71–72. (b)

Masonis, E.J. Comparing two patterns of instruction for teaching item-writing theory and skills. Unpublished doctoral dissertation, University of Pennsylvania, 1971.

McMorris, R.F., Brown, J.A., Snyder, G.W., & Pruzek, R.M. Effects of violating item-construction principles. *Journal of Educational Measurement*, 1972, *9*, 287–295.

Morrison, H.C. *The practice of teaching in the secondary school*. Chicago: University of Chicago Press, 1926.

Nitko, A.J. Criterion-referenced testing in the context of instruction. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Parkhurst, H.H. *Education on the Dalton Plan*. New York: Dutton, 1922.

Pyrczak, F. Objective evaluation of the quality of multiple-choice test items. Unpublished doctoral dissertation, University of Pennsylvania, 1972.

Scandura, J.M., & Durnin, J.H. Assessing behavior potential: Adequacy of basic theoretical assumptions. Philadelphia: University of Pennsylvania, 1971. (mimeo)

Terranova, C. The effects of negative stems in multiple-choice test items. Unpublished doctoral dissertation, State University of New York at Buffalo, 1969.

Villegas, C. Confidence region for a linear relation. *Annals of Mathematical Statistics*, 1964, *35*, 780–788.

Washburne, C.W. *Adjusting the school to the child: Practical first steps*. Yonkers, N.Y.: World Book, 1932.

Williamson, M.L., & Hopkins, K.D. The use of "none-of-these" versus homogeneous alternatives in multiple-choice tests: Experimental reliability and validity comparisons. *Journal of Educational Measurement*, 1967, *4*, 53–58.

PRESCRIBING TEST LENGTH FOR CRITERION-REFERENCED MEASUREMENT

Melvin R. Novick
The American College Testing Program
and
The University of Iowa
and
Charles Lewis
The University of Illinois

In a program of Individually Prescribed Instruction (IPI), where a student's progress through each level of a program of study is governed by his performance on a test dealing with individual behavioral objectives, there is considerable value in keeping the number of items on each test at a minimum. The specified test length for each objective must, however, be adequate to provide sufficient information regarding the student's degree of mastery of the behavioral objective being tested. The minimum acceptable length depends on the manner in which test information is used to make decisions about individual students, the level of functioning required for defining mastery of an objective, the relative losses incurred in making false positive and false negative decisions, the background information available on the student and on the instructional process, and the premium on testing time within the instructional process. Our purpose in this paper is to discuss these issues and provide some broad guidelines for test-length specification for IPI posttests. These specifications will be tentative because of unresolved substantive and methodological issues, but we believe that they will provide some improvement on current practice. A separate, and rather more complex treatment will be required for placement and pretest-length specification.

BACKGROUND

In a criterion-referenced measurement approach to Individually Prescribed Instruction, we imagine a population of test items, having mixed item difficulty, dealing with a particular objective and an ideal decision which advances a student past this objective if he is able to answer at least a given percentage of the items in the population. This minimum

---

The research reported herein was performed pursuant to Grant No. OEG-0-72-0711 with the Office of Education, U.S. Department of Health, Education, and Welfare. Points of view or opinions stated do not, therefore, necessarily represent official USOE position or policy. We are grateful to Charles Davis and Nancy Petersen for helpful comments and computations.

Table 1. Percent of Students Expected To Be Incorrectly Advanced or Retained

Specified Criterion Level .70

Advancement Score	No. of Test Items	Student's True Level of Functioning*									
		50	55	60	65	70	75	80	85	90	95
6	7	6	10	16	23	67	55	42	28	15	4
6	8	15	22	32	43	45	32	20	11	4	1
7	9	9	15	23	34	54	40	26	14	5	1
7	10	17	27	38	51	35	22	12	5	1	-
8	11	11	19	30	43	43	29	16	7	2	-
9	12	7	13	23	35	51	35	20	9	3	-
10	13	5	9	17	28	58	42	25	12	3	-
11	14	3	6	12	22	64	48	30	15	4	-
12	15	2	4	9	17	70	54	35	18	6	-

Specified Criterion Level .75

Advancement Score	No. of Test Items	Student's True Level of Functioning*									
		50	55	60	65	70	75	80	85	90	95
6	8	15	22	32	43	55	32	20	11	4	1
7	9	9	15	23	34	46	40	26	14	5	1
8	10	6	10	17	26	38	47	32	18	7	1
9	11	3	7	12	20	31	55	38	22	9	2
9	12	7	13	23	35	49	35	20	9	3	-
16	20	1	2	5	12	24	58	37	17	4	-
17	21	-	1	4	9	20	63	41	20	5	-
18	22	-	1	3	7	17	68	46	23	6	-

Specified Criterion Level .80

Advancement Score	No. of Test Items	Student's True Level of Functioning*									
		50	55	60	65	70	75	80	85	90	95
6	7	6	10	16	23	33	45	42	28	15	4
7	8	4	7	11	17	26	37	50	34	19	6
8	9	2	4	7	12	20	30	56	40	23	7
8	10	6	10	17	26	38	53	32	18	7	1
9	11	3	7	12	20	31	46	38	22	9	2
10	12	2	4	8	15	25	39	44	26	11	2
11	13	1	3	6	11	20	33	50	31	13	2
12	15	2	4	9	17	30	46	35	18	6	-
17	20	-	1	2	4	11	23	59	35	13	2
19	22	-	-	1	3	7	16	67	42	17	2

Table 1 ***(Continued)***

Specified Criterion Level .85

Advancement Score	No. of Test Items	Student's True Level of Functioning*									
		50	55	60	65	70	75	80	85	90	95
7	8	4	7	11	17	26	37	50	34	19	6
8	9	2	4	7	12	20	30	44	40	23	7
9	10	1	2	5	9	15	24	38	46	26	9
10	11	1	1	3	6	11	20	32	51	30	10
11	12	-	1	2	4	9	16	28	56	34	12
17	19	-	-	1	2	5	11	24	56	29	7
19	21	-	-	-	1	3	8	18	63	35	8

*The true level of functioning is the percent of items a student would be able to answer correctly if he were given the entire universe of items.

Students having true level of functioning values less than the specified criterion level should fail a test composed of all items from this universe. However, on any given test of finite length, some of these students will get more than the minimum advancement percent of the items correct and be considered as "passers." The expected percent of such incorrect advancements is given in the body of the table to the left of the broken line.

Students having true level of functioning values equal to or greater than the minimum advancement percent should pass such a test. The percent of these students who will be incorrectly retained are shown in the table to the right of the broken line.

passing percentage, the so-called *criterion level*, simply reflects the degree of mastery deemed sufficient for this objective (although it implicitly involves the difficulty of the items as well). The actual percentage of items that a student would answer correctly in the population of items is called his *level of functioning*. In practice, the advancement-retention decision must be made from a small sample of observations (test items) and, hence, errors in the decision process must be expected.

One common treatment of the test-length problem in a criterion-referenced measurement context has been given by Millman (1972). He studied a standard decision rule which advances the student if the percent of items correctly answered on a test equals or exceeds the required criterion level. Here it is assumed that the items on the test may be treated as a random sample from the population of interest, so that the obtained percentage correct is a useful estimate of the true population percentage for the student. Using binomial probability tables, Millman obtained the probability that a student with a given *true* level of functioning would be incorrectly advanced or retained by this procedure.

Table 1 expands on some of Millman's computations and gives the conditional probability of incorrect advancement or retention for a variety of true levels, test lengths, and minimum passing percentages. The first impression conveyed by this table is that a substantial proportion (sometimes more than half) of the students with true levels close to, or at the criterion level, will be incorrectly advanced or retained, at least for the test lengths considered. There appears to be a slight improvement in accuracy of decision as the test length increases from 8 to 22 items, although this effect is largely hidden by fluctuation in the probabilities due to changes in the percentage correct required for advancement. For example, with a criterion level of .7, the percentage correct required for advancement is .75, .78, .70, .73, or .75 for test lengths of 8, 9, 10, 11, or 12 items, respectively. This raises a question as to the optimality of the decision procedure assumed in Table 1. To provide a framework for answering this question, let us consider some of the issues involved.

Suppose seven out of eight were taken as the *minimum advancement score* when the criterion level is .75; the probability of incorrect advancement would decrease substantially for all students with true levels below the criterion level. This is shown in Table 2. Those students above .75, on the other hand, suffer a substantial increase in their chances of being incorrectly retained. A more general framework is apparently required before the decision procedure can be chosen, or any judgment be made concerning minimum test length. This framework would need to take into account on which side of .75 small expected errors were considered to be more important.

Table 2. Percent of Students Expected To Be Incorrectly Advanced or Retained

Criterion Level = .75 Test Length = 8

Advancement Score	**True Level**									
	50	55	60	65	70	75	80	85	90	95
6	15	22	32	43	55	32	20	11	4	1
7	4	7	11	11	26	63	50	34	19	6

A FRAMEWORK FOR SPECIFYING TEST LENGTH

Table 1 helps identify the seriousness of the problem of short tests. From a practical standpoint, however, a solution to the problem must involve looking at a different conditional probability and abandoning the simple decision procedure that Millman has so convincingly demonstrated to be inadequate. Instead of the probability that a student will attain a particular test score, given his true level, what is required in making a decision is the probability that a student's true level of functioning exceeds the specified criterion level, given his test score. In other words, it is the test score—not the true level—which is given (i.e. observed) and which is the basis for any decision to advance or retain the student. A student should thus be advanced only if this is sufficiently high probability that he has attained or surpassed the criterion level, *given his test score*. To obtain the necessary probability an application of Bayes' theorem is required. In such an analysis prior knowledge (expressed in probabilistic terms) of the student's true level of functioning is combined with the (binomial) model information relating the observed test score to true level; the result is a posterior probability distribution for true level of functioning, given the test score. The probability this distribution assigns to levels above the criterion is the quantity of interest. In this formulation the problem can be described as selecting a minimum sample size and an advancement score, so that students attaining that score will then have a sufficiently high probability of having at least the minimum required level of functioning.

As a first approximation, let us suppose that prior to having a student's test results our knowledge of his true level of functioning is vague. If this state of knowledge is characterized by selecting a uniform distribution on the interval from zero to unity for true level, π, Bayes' theorem provides the posterior probabilities listed in Table 3 for various scores and test lengths. The posterior distributions on which these probabilities are based all belong to the Beta family, and the parameters in each case are those given in the table, primarily for future reference.

To generate a decision procedure on the basis of Table 3, we must select a criterion level (π_0) and a minimum acceptable probability that a student's true level (π) exceeds this criterion. Thus, for example, we

Table 3. Probability Student's True Level of Functioning Is Greater Than π_0 Given a Uniform Prior Distribution

Minimum Advancement Score	No. of Test Items	Posterior Distribution	Criterion Level—π_0									
			50	55	60	65	70	75	80	85	90	95
6	8	$\beta(7, 3)$	91	85	77	66	54	40	26	14	5	1
7	8	$\beta(8, 2)$	98	96	93	88	80	70	56	40	23	7
8	8	$\beta(9, 1)$	100	100	99	98	96	92	87	77	61	37
7	9	$\beta(8, 3)$	95	90	83	74	62	47	32	18	7	1
8	9	$\beta(9, 2)$	99	98	95	91	85	76	62	46	26	9
9	9	$\beta(10, 1)$	100	100	99	99	97	94	89	80	65	40
7	10	$\beta(8, 4)$	89	81	70	57	43	29	16	7	2	-
8	10	$\beta(9, 3)$	97	93	88	80	69	54	38	22	9	2
9	10	$\beta(10, 2)$	99	99	97	94	89	80	68	51	30	10
8	11	$\beta(9, 4)$	93	87	77	65	51	35	21	9	3	-
9	11	$\beta(10, 3)$	98	96	92	85	75	61	44	26	11	2
10	11	$\beta(11, 2)$	100	99	98	96	92	84	73	56	34	12
9	12	$\beta(10, 4)$	95	91	83	72	58	42	25	12	3	-
10	12	$\beta(11, 3)$	99	97	94	89	80	67	50	31	13	2
11	12	$\beta(12, 2)$	100	100	99	97	94	87	77	60	38	14

might take $\pi_0 = .80$ and the minimum acceptable $\text{Prob}(\pi \geq \pi_0|x, n) = .50$, where x is test score and n is test length. We would then be saying that we wanted to advance the student only if we were at least 50% sure that his level of functioning was above .80. Then, using Table 3, we see that with $n = 8$, all students having $x \geq 7$ would advance to the next objective, but not those with $x = 6$. For a test of 12 items the minimum advancement score would be 10 correct.

Note, however, that if we required 80% assurance that the true level of functioning was above .80, $[\text{Prob}(\pi \geq .80) \geq .80]$, then even those students with eleven correct responses to twelve items would not be advanced. We think that it is unreasonable to require perfect performance as a standard for advancement and, therefore, we need to improve upon this analysis. One way is to use a longer test, but we can at least hope to find a procedure in which a twelve-item test will be adequate.

Although the results in Table 3 provide relevant information for mastery decisions about students based on test scores, they do not take full advantage of the power which is available through the use of prior knowledge. In particular, it will seldom be the case that our knowledge of a student's true level is adequately described by a uniform distribution. For example, our prior probability that a student is functioning above a criterion level of .8 might be approximately .75. This would be the case if historical data suggested that about 75% of the students who completed a unit of Individually Prescribed Instruction proved to be at or above mastery level. Moreover, we might judge the strength of our knowledge to be roughly equivalent to that based on a score from a twelve-item test. (A method for making this assessment will be referred to shortly.)

When working with a binomial model it is convenient and generally very satisfactory to select a member of the Beta class of distributions to characterize prior beliefs (Novick & Jackson, 1974). If this is done the posterior distribution is easily obtained, and in every instance will again be a member of the Beta family. In fact, if the prior distribution is $\beta(a, b)$ and x success in n trials are observed, then the posterior distribution is $\beta(x + a, n - x + b)$. This can be seen in Table 3 where it is noted that the uniform distribution is $\beta(1, 1)$. If we restrict ourselves to prior distributions in the Beta family, the beliefs specified in the previous paragraph are characterized by $\beta(10.254, 1.746)$. Given this prior distribution and the indicated test results, the posterior distributions and posterior probabilities of exceeding various criteria are provided in Table 4. The precise stipulation of prior distributions must always be done carefully, but extensive aids (Novick & Jackson, 1974; Novick, Lewis, & Jackson, 1973) are available, and an elaborate system of Computer Assisted Data Analysis (CADA) is available (Novick, 1973) to help an instructional decision maker specify his prior distribution. An even more sophisticated way of getting prior and posterior distributions for each person is derived by Lewis, Wang, and Novick (1973) and the required tables are given by

Table 4. Probability Student's True Level of Functioning Is Greater Than π_0 Given A $\beta(10.254, 1.746)$ Prior Distribution

Minimum Advancement Score	No. of Test Items	Posterior Distribution	Criterion Level—π_0									
			50	55	60	65	70	75	80	85	90	95
6	8	$\beta(16.254, 3.746)$	100	100	98	96	90	78	60	37	15	2
7	8	$\beta(17.254, 2.746)$	100	100	100	99	97	92	81	62	36	10
8	8	$\beta(18.254, 1.746)$	100	100	100	100	99	98	94	85	66	32
7	9	$\beta(17.254, 3.746)$	100	100	99	97	92	82	65	41	17	2
8	9	$\beta(18.254, 2.746)$	100	100	100	99	98	93	84	66	39	11
9	9	$\beta(19.254, 1.746)$	100	100	100	100	100	98	95	87	69	34
7	10	$\beta(17.254, 4.746)$	100	99	97	93	84	68	47	24	7	1
8	10	$\beta(18.254, 3.746)$	100	100	99	98	93	84	68	45	19	3
9	10	$\beta(19.254, 2.746)$	100	100	100	99	98	95	86	69	42	12
8	11	$\beta(18.254, 4.746)$	100	99	98	94	87	72	51	27	8	1
9	11	$\beta(19.254, 3.746)$	100	100	100	98	95	87	72	48	22	3
10	11	$\beta(20.254, 2.746)$	100	100	100	100	99	96	88	72	45	13
9	12	$\beta(19.254, 4.746)$	100	100	99	96	89	76	55	30	10	1
10	12	$\beta(20.254, 3.746)$	100	100	100	99	96	89	75	52	24	4
11	12	$\beta(21.254, 2.746)$	100	100	100	100	99	96	90	75	48	14

Note: The mean and mode, respectively of $\beta(10.254, 1.746)$ are .855 and .925 and for this distribution Prob $(\pi > \pi_0)$ for π_0 = .70, .75, .80, .85 are .92, .86, .75, and .59, respectively. A close look at these distributional characteristics will help a decision maker determine if this prior distribution is a realistic characterization of his beliefs.

Wang (1973). For the present, we will suppose that this work has been done carefully and that the prior distribution used in the construction of Table 4 is appropriate.

Tables 3 and 4 clearly demonstrate the impact of prior knowledge on our interpretation of test results. In Table 3, for example, the posterior probability that a student with a score of six out of eight items correct has a true level greater than .80 is only .26, whereas in Table 4 this probability has increased to .60. This result should not be surprising in view of the fact that we have now set this probability to be .75 a priori, as compared to .20 in Table 3. If we felt the chances to be very good that the student had mastered an objective (to a level above .8) before we saw the test results, then a score of six out of eight will not substantially change our beliefs; it will lower the probability, but may still leave the odds, a posteriori, in favor of mastery. In many applications a prior probability of mastery may be no more than .60, but the results will still differ sharply from those obtained, assuming vague prior information. Note that if we were to adopt the rule that we will advance a student if the a posteriori probability of mastery is at least .50, then in this example we will advance him if the prior distribution were that of Table 4, but not if it were that of Table 3.

When the decision maker specifies an informative prior distribution he is saying, in effect, that he wants a decision which will have a high probability of being correct in that portion of the decision space in which he thinks the student's ability truly lies. For example, referring to Table 2, a decision maker with a high prior probability that the student had a true level of functioning below .75 would, by virtue of his analysis, require a minimum passing score of seven correct out of eight items. This would assure him a low probability of misclassification for all values below .75. Another decision maker with high prior probability that the student was above criterion level would likely require only six out of eight correct, and thus have low probability of an incorrect decision for values of .75 or above.

Once we have decided to work with the posterior probability that a student's level of functioning exceeds some criterion, given his test score, and have made use of our prior knowledge in obtaining this probability, another issue remains to be settled before we can turn to the question of test length. Simply stated, we need to know how sure we should be that a student has mastered an objective at the chosen level before we make the decision to allow him to advance to the next objective. For instance, is a posterior probability of at least .5, as was used in the last example, a reasonable choice in all cases? Almost certainly this last question should be answered in the negative. The point at issue here comes down to an understanding of the relative disutilities or losses associated with the false positive and false negative errors.

If it were no more serious to advance a student whose level was below the criterion than to retain a student who was above, we would be behaving optimally if we were to advance students with posterior probabilities above .5 and retain the others. In many situations the prior probability will be this high, and hence an advancement decision could then be made on an a priori basis. On the other hand, we might consider the loss to be twice as great for a false advancement than for a false retention. In this case we should advance only those students whose posterior probability for being above the criterion exceeds ⅔. The general result is that we will achieve the smallest expected loss if we match the posterior odds to the loss ratio. Thus, if the loss ratio is 2 to 1 (false advance to false retain), a probability of 2/(2 + 1) gives matching odds of ⅔ to ⅓ above criterion to below criterion.

Table 5. Losses Associates With Incorrect Decisions

		True Level	
		$\pi \geq \pi_0$	$\pi < \pi_0$
Decision	Advance	0	a
	Retain	b	0

To express the result symbolically, consider the notation of Table 5. Here *a* is the loss associated with advancing a student whose true level is below π_0, and *b* is the loss for retaining a student whose true level exceeds π_0. The decision rule which minimizes expected loss in this situation is to advance a student if his test score is such that b $\text{Prob}(\pi \geq \pi_0|x, n) \geq$ a $\text{Prob}(\pi < \pi_0|x, n)$, and to retain him otherwise. This comparison is equivalent to comparing the loss ratio a/b to the probability ratio $\text{Prob}(\pi \geq \pi_0|x, n) / \text{Prob}(\pi < \pi_0|x, n)$.

If a = b in our analysis the decision procedure reduces to comparing the median of the posterior distribution with the specified criterion level. If the median is at least at this level the student is advanced, otherwise he is retained. In this situation the decision procedure is very similar to that used by Millman (1972). Though the procedure used by Millman is not Bayesian, it is equivalent to comparing with the mode (rather than the median) of the posterior distribution based on a uniform prior. Thus, in effect, the sampling theory approach gives equal weight to all equal intervals throughout the range of π; that is, effectively, to take π to be uniformly distributed a priori. This is seldom a reasonable prior specification. We might also remark that the formulation in Table 5 can be generalized to provide for differential *utilities* for correctly identifying true positives and true negatives as well as differential disutilities (or losses) for false positives and false negatives as is done in Table 5. To do this, negative quantities (negative disutilities = utilities) would need to replace the zeros in Table 5, and a slightly more complicated analysis would be used.

It may be worthwhile to summarize the situation at this point. An instructor wishing to use test results in the context of Individually Prescribed Instruction should be ready to supply three kinds of information. First, a criterion level—the minimum degree of mastery required—must be set. In Individually Prescribed Instruction this seems to run from about .70 to about .85. Second, prior knowledge of the student's true level of functioning must be translated into probability terms, namely a prior probability distribution for π. Typically, a carefully monitored program will be such as to suggest a prior probability distribution that assigns a probability of just more than .50 to the region above the criterion level. If this is not the case, the general efficacy of the program should be re-evaluated. A program that results in a much higher probability may be wastefully long and one that results in a lower probability may require strengthening. Finally, the relative losses associated with the two types of incorrect decisions must be assessed. A ratio of more than 1/1 is the rule (we are told) with ratios of 1.5/1 and 2/1 being common, and ratios as high as 3/1 not being rare.

It should be clear that all three of the above determinations will have a bearing on the minimum necessary test length. As the criterion level approaches unity the test must be longer in order to provide adequate information about a student's level of functioning in the neighborhood of the criterion. If prior probabilities of mastery are sufficiently high, very short tests become possible, but this is not and should not be the typical case. Finally, higher loss ratios require longer tests to allow the possibility of high posterior probability of mastery. We shall also see that greater test lengths are sometimes required because of the obvious restriction to integer valued sample sizes.

A DESIGN FOR TEST-LENGTH SPECIFICATION

The characteristics of the *group* of students being tested must now be considered as they relate to test-length specification. Each member of the group of students tested has been exposed to the same instructional program under identical local conditions. If a particular student is not considered atypical for this group, then our prior beliefs about his true level of functioning should closely reflect the true distribution of levels of functioning found in that group. Indeed, elaborate formal procedures for effectively bootstrapping a prior distribution using, for each examinee, the scores on the remaining m – 1 examinees are described by Novick, Lewis, and Jackson (1973). Group characteristics, thus, through their effect on our prior distributions, do affect test-length specification. If the average test score of the group is high (i.e., above the criterion level) and there is little variation among individuals, shorter tests become feasible.

In practice, since prior distributions will be based upon on-site experience, there will of course be different prior distributions for different sites. What we will attempt to do here is to show what sample sizes will be required for a broad range of prior distributions and loss ratios. What we need to do now, therefore, is to consider certain combinations of prior distributions, criterion levels, and loss ratios, and see what sample size will be adequate in each case.

For our analyses we will consider 20 different prior distributions for the level of functioning π, four specified criterion levels, and four loss ratios. For each criterion level we will consider all four loss ratios and four of the prior distributions. The four loss ratios we will use are 1.5, 2.0, 2.5, and 3.0. The respective probabilities $P = \text{Prob}(\pi > \pi_o)$ required for advancement [given by setting $P/(1\text{-}P)$ equal to the loss ratios, a/b] are .60, .67, .71, and .75. Thus, with a loss ratio of 3.0, the posterior probability that the student's level of functioning is greater than the specified criterion level must be at least .75, if he is to be advanced.

The 20 prior probability distributions we will be considering are given in Table 6 where they have been grouped in blocks of five, with each block having a distribution with the respective mean values .70, .75, .80, .85, and .90. The blocks differ with respect to the concentration of the prior distributions. Within block the distributions differ with respect to their mean values. Note that in the first block the arguments of each Beta distribution sum to 8, e.g., $5.6 + 2.4 = 8$. This indicates that the amount of prior information contained in each of these distributions is equivalent to what would be gained from a test containing eight items. Given one of these prior distributions and some criterion level and loss ratio, if we specify an eight-item test our posterior distribution will contain information equivalent to that contained in 16 observations. This contrasts with the classical procedure which uses no prior information. It is this increment in information that is equivalent to prior observations which permits a reduction in test length when a Bayesian procedure is used.

The first problem in doing an analysis is that of selecting a reasonable prior distribution. For the present application we would first need to ask ourselves what we would expect to find as the mean level of functioning in our posttest group. With a specified criterion level of .70 we might hope for a mean level of functioning of .70. Thus, we would have people in training until such time as we would "expect" them to be qualified. Since loss ratios are typically greater than one, some overtraining may be thought to be useful but, as we shall see, excessive overtraining may be wasteful.

Suppose, for concreteness, that we believe the mean population level of functioning to be .70. Distributions 1, 6, 11, and 16 satisfy this condition and, hence, we may choose from among them. We note that these distributions are in an increasing order of tightness, as may most conveniently be seen in the probability assignment given in the last column, to the interval (.90, 1.00). These probabilities are respectively .08, .05, .03, and

Table 6. Selected Prior Distributions for IPI Advancement Decisions

No.	Prior Distribution	Effective Prior Sample Size	Mean	Prob ($\pi_l \leq \pi \leq \pi_u$)* .00-.70	.70-.75	.75-.80	.80-.85	.85-.90	.90-1.00
1	$\beta(5.6, 2.4)$	8	.70	.46	.12	.12	.12	.10	.08
2	$\beta(6,2)$	8	.75	.33	.12	.13	.14	.13	.15
3	$\beta(6.4, 1.6)$	8	.80	.21	.10	.12	.15	.16	.26
4	$\beta(6.8, 1.2)$	8	.85	.12	.07	.09	.13	.17	.42
5	$\beta(7.2, .8)$	8	.90	.05	.04	.06	.09	.14	.62
6	$\beta(7, 3)$	10	.70	.46	.14	.14	.12	.09	.05
7	$\beta(7.5, 2.5)$	10	.75	.32	.13	.15	.15	.13	.12
8	$\beta(8, 2)$	10	.80	.20	.10	.14	.16	.17	.23
9	$\beta(8.5, 1.5)$	10	.85	.10	.07	.10	.14	.19	.40
10	$\beta(9, 1)$	10	.90	.04	.03	.06	.10	.16	.61
11	$\beta(8.4, 3.6)$	12	.70	.47	.15	.15	.12	.08	.03
12	$\beta(9, 3)$	12	.75	.32	.14	.16	.16	.13	.09
13	$\beta(9.6, 2.4)$	12	.80	.18	.11	.15	.18	.18	.20
14	$\beta(10.2, 1.8)$	12	.85	.09	.07	.11	.16	.20	.37
15	$\beta(10.8, 1.2)$	12	.90	.03	.03	.06	.11	.17	.60
16	$\beta(10.5, 4.5)$	15	.70	.47	.17	.16	.12	.06	.02
17	$\beta(11.25, 3.75)$	15	.75	.30	.16	.18	.17	.13	.06
18	$\beta(12,3)$	15	.80	.16	.12	.17	.20	.19	.16
19	$\beta(12.75, 2.25)$	15	.85	.07	.07	.12	.18	.23	.33
20	$\beta(13.5, 1.5)$	15	.90	.02	.03	.06	.11	.19	.59

*Note: All entries have been rounded to two decimal places and smoothed so that the row totals add to 1.00.

.02. We need to ask ourselves which of these values seems most reasonable, and this then will give us some preference among these prior distributions. We might consider the relative weight of prior information assumed by each prior distribution (8, 10, 12, and 15 equivalent prior observations, respectively), and this should help to narrow our focus to one or two adjacent prior distributions for this, or any other application. Since we cannot know what an appropriate prior distribution will be in applications we have not seen, it will be most helpful, we think, to work out sample-size allocations for several prior distributions and leave the final selection to be made "in the field." We believe that the prior distributions, loss ratios, and specified criterion levels used here are typical of those found in practice and, therefore, that the specific results we will obtain will be useful. However, if other combinations present themselves, we believe that the general methodology that we are demonstrating should be adequate to the problem. Actually we shall find that most of our specifications are very robust with respect to the choice of prior distribution within the range we have considered.

SOME SPECIFIC TEST-LENGTH RECOMMENDATIONS

In Table 7 we give recommended sample sizes and minimum advancement scores for $\pi_0 = .70$, (a/b) = 1.5, 2.0, 2.5, 3.0 and prior distributions 1, 6, 11, and 16. The values that we have settled on for the body of this table are not, in every instance, optimum in any statistical sense, though we are confident that the risks associated with these decision rules are in every case insignificantly different from the risks of the optimum procedures. In selecting values for this table we have sought sample sizes and minimum advancement scores that would be very efficient over a wide range of prior distributions. That we have been successful in this

Table 7. Recommended Sample Sizes and Advancement Scores

		$\pi_0 = .70$			
		Loss Ratio			
Prior Distribution	$\mathcal{E}(\pi)$	1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
$\beta(5.6, 2.4)$[1]	(.70)	6/8(.62)	10/13(.70)	11/14(.74)	12/15(.78)
$\beta(7, 3)$	(.70)	6/8(.61)	10/13(.69)	11/14(.73)	12/15(.77)
$\beta(8.4, 3.6)$	(.70)	6/8(.61)	10/13(.68)	11/14(.72)	12/15(.76)
$\beta(10.5, 4.5)$	(.70)	9/12(.62)[2]	10/13(.67)	11/14(.71)	12/15(.75)
		General Recommendations			
		6/8(75%)	10/13(77%)	11/14(79%)	12/15(80%)

[1] Apriori, Prob($\pi \geq .70$) for each of the four prior distributions is .54, .54, .53, and .53.

[2] For 6/8, Prob($\pi \geq .70$) = .598.

endeavor is confirmed by our ability to give general recommendations that hold throughout the range of prior distributions studied. In only one instance have we actually cheated (see footnote 2, Table 7), but again the increase in expected loss will be trivial. We would also note that the required percentage correct and the number of required observations increase as the loss ratio increases, which "makes sense" on intuitive grounds.

A rough indication of the near optimality of any of the individual specifications can be gained from the closeness of the a posteriori probability (indicated in parentheses following the specification) to the value required by the particular loss ratio (given in parentheses at the top of the column). Thus, with the prior distribution $\beta(7, 3)$, the decision rule "six out of eight," abbreviated 6/8, leads to the a posteriori distribution $\beta(13, 5)$ and to $\text{Prob}(\pi > .70) = .61$ which is just .01 greater than the required level .60 for the loss ratio 1.5 (1.5 to 1). In this instance, the specified decision rule may be very good. On the other hand, consider the prior distribution $\beta(5.6, 2.4)$. Here the rule 11/14 leads to a value .74 when only .71 is required for a 2.5 to 1 loss ratio. Actually, the specification 8/10 is somewhat better giving a posterior probability of .729. Also for the prior distribution $\beta(7, 3)$, the posterior probability with 8/10 is .718. With the loss ratio 2.0/1 and with the prior $\beta(5.6, 2.4)$, the rule 7/9 leads to the posterior probability .68 as compared to desired value of .67. In every case where we have specified an "almost best" decision rule, the result has been an increase in the specified sample size and the purpose has been to obtain uniformity of specification over a reasonably wide range of amounts of prior information. Considering our general ignorance concerning what might be an appropriate prior distribution in specific applications, the specifications we have given should be the more generally useful.

Another indication of how good a particular specification is can be inferred from the closeness of the percentage correct required by the advancement rule to the specified criterion level. Clearly, if the percentage required by the advancement rule is very much larger than the specified criterion level, a large percentage of qualified students will be retained and this is undesirable, particularly for small loss ratios. For large loss ratios this is less important and hence higher advancement ratios can, and will need to be tolerated. This feature is exhibited in Table 7, where the advancement ratios increase with increasing loss ratios. One can, of course, keep the advancement ratio down very close to the specified criterion level even for higher loss ratios, but only by having much larger sample sizes. For example, with the prior distribution $\beta(5.6, 2.4)$, the specified criterion level $\pi_0 = .70$, and the loss ratio 2.0, the advancement ratio 72/100 is satisfactory since $\text{Prob}(\pi > .70 | 72/100) = .675$, but the indicated sample size is unacceptable.

Note that for each of the prior probabilities used in Table 7,

Table 8. Recommended Sample Sizes and Advancement Scores

		$\pi_0 = .75$			
		Loss Ratio			
Prior Distribution	$\mathcal{E}(\pi)$	1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
$\beta(6, 2)$[1]	(.75)	8/10(.65)	16/20(.70)	17/21(.74)	18/22(.77)
$\beta(7.5, 2.5)$	(.75)	8/10(.64)	16/20(.69)	17/21(.73)	18/22(.76)
$\beta(9, 3)$	(.75)	8/10(.63)	16/20(.69)	17/21(.72)	18/22(.75)
$\beta(11.25, 3.75)$	(.75)	8/10(.62)	16/20(.68)	17/21(.71)	19/23(.77)[2]
		General Recommendations			
		8/10(80%)	16/20(80%)	17/21(81%)	18/22(82%)

[1] Apriori, Prob($\pi \geq .75$) = .56, .55, .55, and .54, respectively, for the four prior distributions used in Table 8.

[2] For 18/22, Prob($\pi \geq .75$) = .744.

Prob($\pi \geq .70$) > .50. On an a priori basis, therefore, advancement would be indicated with a loss ratio 1.0. This will generally be true for the prior distributions we will be adopting for our analyses. The point is that loss ratios of 1.0 are not (we are told) typical of IPI applications, and if test lengths are to be kept reasonable it will be necessary to use training programs that give *mean* output at or above the criterion level.

There has been a definite tendency in IPI to require relatively high advancement ratios; typically, the value .85 is used. One might well ask whether this is a function of a high loss ratio combined with a desire for a short test length, or whether it really reflects a perceived need for a high criterion level. (For example, an advancement ratio of 6/7 with the prior distribution $\beta(5.6, 2.4)$ would yield with $x = 6$ a posterior Prob($\pi > .70$) = .77 which would be just right with a loss ratio of 3.0.) We do not know the answer to this question, but hope that those within IPI will want to consider it carefully. Only through such serious consideration can the test-length problem be "solved."

Some recommended test lengths for $\pi_0 = .75$ and four prior distributions with $\mathcal{E}(\pi) = .75$ are given in Table 8. Again we have been able to specify one generally satisfactory advancement ratio for each of the four loss ratios. We note that the required test lengths for $\pi_0 = .75$ are rather larger than for $\pi_0 = .70$. In Table 8, we find very short required test lengths for a 1.5 loss ratio and rather long ones for loss ratios of 2.0, 2.5, and 3.0.

In Table 9 we provide recommendations for $\pi_0 = .80$ when $\mathcal{E}(\pi) = .80$. The results here parallel those of Table 8, except that the advancement ratios are very high as compared to the criterion levels. This is relatively unsatisfactory. In Footnote 1 to Table 9 we indicate the formal results for the prior distribution $\beta(6.4, 1.6)$ and the sample result "8.5" correct

Table 9. Recommended Sample Sizes and Advancement Scores

		$\pi_0 = .80$			
		Loss Ratio			
Prior Distribution	$\mathcal{E}(\pi)$	1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
$\beta(6.4, 1.6)$[1]	(.80)	6/7(.66)	7/8(.70)	17/20(.72)	19/22(.78)
$\beta(8,2)$	(.80)	6/7(.65)	7/8(.69)	17/20(.72)	19/22(.77)
$\beta(9.6, 2.4)$	(.80)	6/7(.64)	7/8(.68)	17/20(.71)	19/22(.76)
$\beta(12, 3)$	(.80)	6/7(.63)	7/8(.67)	18/21(.73)[2]	19/22(.75)
		General Recommendations			
		6/7(86%)	7/8(88%)	17/20(85%)	19/22(86%)

[1] Apriori, Prob($\pi \geq .80$) = .57; for 8/10, Prob($\pi \geq .80$ = .55; for 16/20, Prob ($\pi \geq .80$) = .54; for 8.5/10, Prob($\pi \geq .80$) = .67; for 8.3/10, Prob($\pi \geq .80$) = .62; for 9/10, Prob($\pi \geq .80$) = .78.

[2] For 17/20, Prob($\pi \geq .80$) = .70.

and "1.5" incorrect and also for "8.3" correct and "1.7" incorrect. These provide very nice results for loss ratios of 2.0 and 1.5, respectively. Unfortunately, these are unobtainable sample results. This demonstrates that in part large required test lengths may sometimes be due to the discreteness and, hence, discontinuity of our possible experimental outcomes. This also suggests that the precise specification of the advancement rules may be highly sensitive to the mean value of the prior distribution even if it is proving to be relatively insensitive to the total amount of information contained in the prior distribution, which is indicated by the sum of the two parameters of the Beta distribution.

For example, given the prior distribution $\beta(6.4, 1.6)$ and the impossible sample result $x = 8.3$, $n = 10$, we have the posterior distribution $\beta(14.7, 3.3)$ which, as we indicated previously, gives Prob($\pi > .80$) = .62 which suggests that the advancement ratio 8.3/10 might be very favorable with a loss ratio of 1.5. But suppose we had just a slightly different prior distribution, namely, $\beta(6.7, 1.3)$ with $\mathcal{E}(\pi) = .84$, then the sample result $x = 8$, $n = 10$ would yield the posterior distribution $\beta(14.7, 3.3)$ and thus, for the reasons given above, indicate that the advancement ratio 8/10 might be attractive. This advancement ratio is clearly more attractive than the ratio 6/7, despite the fact that it requires three additional items, because this ratio 8/10 = 80% is closer to the criterion level than is the advancement ratio 6/7 = 86%.

Because of this relatively high dependence of the results on the expected value of the prior distribution, it seems important to attempt some study of the variation of our results as a function of changes in our prior distribution. For this reason, in Table 10 we have redone our sample size recom-

Table 10. Recommended Sample Sizes and Advancement Scores

		π_0 = .80			
		Loss Ratio			
Prior Distribution	$\mathcal{E}(\pi)$	1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
β6.8, 1.2)[5]	(.85)	8/10(.64)	9/11(.69)	10/12(.72)[1]	11/13(.76)
β(8.5, 1.5)	(.85)	8/10(.66)	9/11(.70)	10/12(.73)[2]	11/13(.76)
β(10.2, 1.8)	(.85)	8/10(.67)	9/11(.71)	9/11(.71)[3]	11/13(.77)
β(12.75, 2.25)	(.85)	8/10(.69)	9/11(.72)	9/11(.72)[4]	11/13(.78)
			General Recommendations		
		8/10(80%)	9/11(82%)	10/12(83%)	11/13(85%)

[1] For 5/6, Prob($\pi \geq .80$) = .72.
[2] For 5/6, Prob($\pi \geq .80$) = .73.
[3] For 10/12, Prob($\pi \geq .80$) = .74.
[4] For 10/12, Prob($\pi \geq .80$) = .75.
[5] For the four prior distributions, the apriori probabilities of $\pi \geq .80$ are .72, .73, .74, and .75. With these prior distributions and with 7/10, the posterior probabilities of $\pi \geq .80$ are .41, .43, .46, and .48.

mendations under the assumption that the mean of our prior distribution is .85 instead of .80.

Surely the practitioner will find the sample size recommendations of Table 10 to be attractive. With these prior distributions, apparently, test lengths need be no greater than 13 for any of the listed loss-ratios. With the prior distributions having $\mathcal{E}(\pi)$ = .80, a sample size of 22 is required when the loss ratio is 3.0.

What is happening is that we are beginning with fairly strong beliefs that $\pi \geq \pi_0$ so that not much data, in confirmation, is required even for high loss ratios. In fact, even on an a priori basis, an advancement decision would be made for all loss ratios up to and including 2.5. Indeed, we see that the function of the sample data here is to provide the possibility of obtaining some information that might change the decision to retention. For example, an observed performance ratio of 10/13 with the prior distribution β(6.8, 1.2) would give a posteriori Prob($\pi \geq .80$) = .72, and hence, the student would be retained if the loss ratio were 3.0 (see also Footnote 5, Table 10).

We believe that the comparison of the specifications in Tables 9 and 10 have important implications for IPI management. When loss ratios are high it may well be advantageous to strengthen the training program to the extent that the mean output is well above the specified criterion level. This will make it possible to use short tests or, alternatively, will generally reduce the risk of incorrect classification. This will, of course, be more expensive, and the investment must be balanced out against the

Table 11. Recommended Sample Sizes and Advancement Scores

	$\pi_0 = .85$				
Prior			**Loss Ratio**		
Distributions	$\mathcal{E}(\pi)$	1.5 (.60)	2.0 (.67)	2.5 (.70)	3.0 (.75)
$\beta(6.8, 1.2)^1$	(.85)	7/8(.62)	9/10(.70)	17/19(.73)	18/20(.76)^3
$\beta(8.5, 1.5)$	(.85)	7/8(.62)	9/10(.69)	17/19(.72)	19/21(.77)
$\beta(10.2, 1.8)$	(.85)	7/8(.61)	9/10(.68)	17/19(.72)	19/21(.76)
$\beta(12.75, 2.25)$	(.85)	7/8(.60)	9/10(.67)	17/19(.71)^2	19/21(.75)
			General Recommendations		
		7/8(87.5%)	9/10(90%)	17/19(89%)	19/21(90%)

[1] The apriori probabilities for $\pi \geq .85$ are .59, .58, .58, and .57.
[2] For 10/11, Prob($\pi > .85$ = .695).
[3] For 19/21, Prob($\pi > .85$ = .78).

reduction in the cost of testing and the reduction in the expected loss due to incorrect decision. The final table, Table 11, looks very much like Table 9 so far as test lengths are concerned. Here again some robust length assignments are obtained, but again the lengths for the high loss ratios border on being discomforting. This can be corrected by training to an average level of functioning of .90. With the prior distribution $\beta(7.2, 8)$, we find that Prob($\pi \geq .85$) = .76 a priori. Observing 6/7 yields Prob($\pi \geq .85$) = .70, while 5/7 yields a value of .41. Observing 8/9 yields .77, while 7/9 yields .493. Clearly, very short test lengths are again possible if the students are trained to a sufficiently high average standard.

SOME SUMMARY REMARKS

The test-length recommendations given in this paper are meant to be taken seriously and hopefully they will soon be adopted on a provisional and experimental basis, so that more experience can be gained while some of the theoretical and substantive issues raised in the paper are debated. The questions of level of functioning required to define mastery and the relative losses incurred in making false positive and false negative decisions require serious discussion. We also need to get some clear picture of what kinds of distributions of outcomes are to be expected as this determines the amount of prior information available in making individual assessments. This third issue is, as we have indicated, intimately related to the expected level of functioning that is sought in the group being t[illegible]. Hopeful and possible outcomes of such discussions could be a consen[illegible]s that:

1. In most situations a level of functioning of something less than .85 is satisfactory. A value as low as .75 would be highly desirable. This could be accomplished by redefining the task domain slightly so as to eliminate very easy items.

2. Training should be carefully monitored so that expected group performance will be just slightly higher than the specified criterion level. This will keep training time and testing time relatively low.
3. The program should be structured so that very high loss ratios are not appropriate. That is, individual modules should not be overly dependent on preceding ones.

One problem that does not arise with Bayesian methods is any complication if sequential methods are used. Items can simply be administered until it is clear that a student will definitely, or cannot possibly, attain the minimum advancement score. Thus with a minimum advancement score of 8/10, testing can cease as soon as eight successes or three failures are observed.

Two important issues that have been treated in a rather gross way in this paper stand in need of further research. It must be recognized that while the threshold loss function we have adopted here is a better approximation to reality than, for example, Livingston's criterion centered squared-error loss (see Hambleton & Novick, 1973), it is only a gross approximation to be used while better and more complicated approximations are being investigated. Three that immediately come to mind are:

1. A threshold loss function with an indifference region in which there is zero loss for false positive or false negative errors.
2. A *negative* squared-exponential loss used with the root arcsine transformation parameter

$$\gamma = \sin^{-1} \sqrt{\pi}$$

3. A cumulative Beta distribution loss function.

We expect that these loss functions will give somewhat different and surely better length specifications than those obtained here, but the overall decrease in expected loss may or may not be great. We should also remark that these recommendations are specifically made for first-time through decisions; we have yet to consider the problem of decisions for students repeating a unit.

Finally, we would remark that one of the important issues that we identified at the outset of this paper has been handled in a most casual and informal manner. To do other than this would have enormously complicated the analysis and delayed substantially the appearance of our recommendations. We refer explicitly to the premium on testing time within the instructional process and implicitly to an implied trade-off between training and testing time. A completely general analysis would consider an available time T and an allocation of T into instruction and testing times i + t = T, so as to maximize a payoff function which would have a (possibly differential) positive payoff for each module successfully completed, and a (differential) negative payoff for an incorrect decision of either type. We are reluctant to undertake such a sophisticated analysis until such time as the operating conditions of IPI are more clearly defined.

For the present paper we have implicitly adopted some guidelines which effectively say that it is very desirable to have test lengths of 12 or less, tolerable but undesirable to have test lengths as high as 20, and discomforting to have tests that are longer than this. We have also taken the position that a decision should not be made on the basis of prior and collateral information alone but that mastery must be confirmed by a test that permits demonstration of nonmastery. As in all of the judgmental decisions made in this paper we have been guided by counsel from experienced IPI personnel, particularly Richard Ferguson and Anthony Nitko to whom we are much indebted. The value of this paper will largely be determined by the quality of the discussion it engenders among such people.

REFERENCES

Hambleton, R., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*(3), 159–170.

Lewis, C., Wang, M., & Novick, M.R. Marginal distributions for the estimation of proportions in m groups. *ACT Technical Bulletin* No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Millman, J. Determining test length, passing scores and test lengths for objectives-based tests. Los Angeles: Instructional Objectives Exchange, 1972.

Novick, M.R. High school attainment: An example of a computer-assisted Bayesian approach to data analysis. *International Statistical Review*, 1973, *41*, 264–271.

Novick, M.R., & Jackson, P.H. *Statistical Methods for Educational and Psychological Research*. New York: McGraw-Hill, 1974.

Novick, M.R., Lewis, C., & Jackson, P.H. The estimation of proportions in m groups. *Psychometrika*, 1973, *38*, 19–46.

Wang, M. Tables of constants for the posterior marginal estimates of proportions in m groups. *ACT Technical Bulletin* No. 14. Iowa City, Iowa: The American College Testing Program, 1973.

EMPIRICAL VALIDATION OF CRITERION-REFERENCED MEASURES

J. Ward Keesling
Center for the Study of Evaluation

In this paper, the term "criterion-referenced measure" indicates a performance task or set of test items which yields a score having a direct, meaningful interpretation. An example of a performance task would be a typing test which yields a score of number of words typed per minute by which one can judge whether or not a particular typist will be able to accomplish a specific piece of work in a given amount of time. A set of test items defined by sampling from a universe of "interchangeable" items (as discussed in Harris' paper on mastery tests) would yield an estimate of the proportion of all such items which a person might be expected to answer correctly. Another type of direct, meaningful interpretation of scores from such tests would be obtained by categorizing the possible outcomes as indicators of mastery or non-mastery.

In this paper, analytical methods for dealing with both the continuous score case and the dichotomous (mastery, non-mastery) case are presented. The tests or performance tasks are presumed to exist already and to have acceptable face validity.

As Harris indicated, the instructional context in which tests are used is very important. It seems to me that the development of systems of behavioral objectives and the application of them in instructional settings entail implicit structures for instruction and testing. Two principal structures can be discerned: (1) Objectives or skills are collected into sets within which there are no a priori notions of order of presentation or of transfer of training. That is, these objectives can be taught and learned in any order whatsoever. The set of objectives used in National Assessment (see the paper by Wilson in this monograph) is an example of this. (2) Objectives are subject to a priori ordering based upon task analysis or theories of instruction. The theory of instruction or the task analysis specifies that certain objectives must be learned before others. In the extreme case, instruction proceeds in the order specified and students are permitted to advance to new objectives only as they master the presumed prerequisites. (The paper by Nitko in this monograph gives further details on this type of organization of instruction.)

In the following discussion, the case of unordered objectives is reviewed briefly. Then the case of ordered objectives is considered with some further distinctions offered within this type of structure.

The first case to be considered is that of the diffuse or unordered set of objectives. As there are no a priori orderings of skills or objectives, there should be no dependencies among the tests or performance tasks associated with these objectives. The notion of instructional level developed earlier by Harris is crucial here. If one group has had more experience with two objectives or skills than another group, mixtures of subjects from both groups will yield data indicative of dependencies between the tasks or tests used as criterion referenced measures. Thus, given a population of subjects with the same amount of experience with two or more skills or objectives, we expect no dependencies among tests or performance tasks used to measure achievement.

In the continuous case this means that the covariance matrix of the measures should be a diagonal. This implies that all pairs of tests or tasks have a correlation of zero. Morrison (1967) gave Bartlett's test of the hypothesis that a covariance matrix is diagonal, which has the form:

$$\chi^2 = -\left(N - 1 - \frac{2p + 5}{6}\right) \ln | R |$$

Where N is the number of subjects, p is the number of measures and ln | R | is the natural logarithm of the determinant of the correlation matrix of the p measures. If the obtained χ^2 is less than the tabled value of χ^2 for the selected level of α and $p(p - 1)/2$ degrees of freedom, we accept the null hypothesis of no linear relationship among the measures.

In the dichotomous case, one could develop a series of two by two tables designed to test the hypothesis that there are no associations between pairs of measures. Naturally, this would lead to a large number of tests with the attendant risk of finding one or more significant associations by chance alone. An overall test of the hypothesis of no pair-wise association could be developed using the techniques presented by Goodman and Kruskal (1972).

The above procedures can be likened to the notion of discriminant validation developed by Campbell and Fiske (1959). Criterion referenced measures of conceptually independent behaviors should not correlate.

Previously, Harris presented a discussion of the use of three types of criterion information in the validation of mastery tests. If, instead only of mastery tests, we also consider the more general case of a criterion referenced measure and if we allow the validity information to be either dichotomous or continuous in nature, then we need to include the standard correlation and regression procedures in our armamentarium of analytic techniques. In all cases we would look for associations between the criterion-referenced measure in question and the criterion information contained in the instructional history or the performance on another task.

RELATED OBJECTIVES

Within the general category of related objectives, we will distinguish three cases: (1) The hierarchical chain of independent, additive components, (2) The complex of ordered, independent, additive components, and (3) The complex of ordered, interrelated additive components.

The first case, that of the hierarchical chain of independent, additive components, may be represented graphically and algebraically as follows:

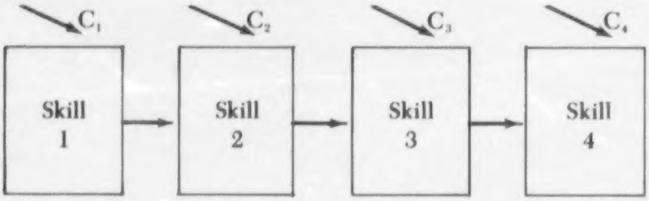


Figure 1: Hierarchical Chain of Independent Additive Components

Skill 2 is composed of skill 1 plus a new component, shown as C_2. This component is acquired independently of skill 1 and enables the subject to perform at the level of complexity of skill 2 where he would otherwise perform only at the level of skill 1. If, as in the Taxonomy of Educational Objectives developed by Bloom, et al. (1956), knowledge (skill 1) must precede comprehension (skill 2), then C_2 is that additional component which, when acquired, permits comprehension as distinct from knowledge.

CONTINUOUS SCORE CASE

The algebraic formulation for a model of independent components such as this has been presented by Guttman (1955):

$$\begin{aligned} t_{j1} &= c_{j1} + e_{j1} \\ t_{j2} &= c_{j1} + c_{j2} + e_{j2} \\ t_{j3} &= c_{j1} + c_{j2} + c_{j3} + e_{j3} \\ t_{j4} &= c_{j1} + c_{j2} + c_{j3} + c_{j4} + e_{j4} \end{aligned} \tag{2}$$

Here, t_{ji} $(i = 1, \ldots, 4)$ is the score of subject j on a test of skill i and the c_{ji} are the independent additive components which combine, in an ordered fashion, to produce the various levels of complexity of the skills. The e_{ji} are errors in the observations. Further specifications on the algebraic model are:

$$\begin{aligned} &\text{Covariance } (c_k\, c_m) = 0 \quad (k \neq m) \\ &\text{Covariance } (e_k\, e_m) = 0 \quad (k \neq m) \\ &\text{Covariance } (e_k\, c_m) = 0 \quad (k, m = 1, \ldots, n) \\ &\text{Covariance } (e_k\, t_m) = 0 \quad (k \neq m) \end{aligned} \tag{3}$$

These specifications indicate that the errors e_{ji} are independent from test to test and that the errors do not correlate with any of the independent "true" component scores. With these specifications we may write out the covariance matrix of the observed test scores, Σ_t, in terms of the variances and covariances of the components, c_{ji}, and the errors, e_{ji}.

$$(4)\quad \begin{bmatrix} \sigma^2_{c1}+\sigma^2_{e1} & & & \text{Symmetric} \\ \sigma^2_{c1} & \sigma^2_{c1}+\sigma^2_{c2}+\sigma^2_{e2} & & \\ \sigma^2_{c1} & \sigma^2_{c1}+\sigma^2_{c2} & \sigma^2_{c1}+\sigma^2_{c2}+\sigma^2_{c3}+\sigma^2_{e3} & \\ \sigma^2_{c1} & \sigma^2_{c1}+\sigma^2_{c3} & \sigma^2_{c1}+\sigma^2_{c2}+\sigma^2_{c3} & \sigma^2_{c1}+\sigma^2_{c2}+\sigma^2_{c3}+\sigma^2_{c4}+\sigma^2_{e4} \end{bmatrix}$$

A covariance matrix such as this may be transformed into a correlation matrix which will have a pattern called the "simplex" by Guttman (1955). The elements in the correlation matrix will all be positive and will diminish in magnitude as one moves away from the diagonal along any row or column. Thus, one easy test of the validity of a set of criterion referenced tests for skills arrayed in a strict hierarchical order would be to compute the correlation matrix of the tests, ordered either from least complex skill to most complex skill or vice-versa, and observe whether or not it has the simplex pattern. In a subsequent section of this paper I present a more formal mathematical model for this case. The reader is referred to articles by Kaiser (1962), Mukherjee (1966), and Schönemann (1970) for additional work on simplex models.

In order to use this technique as a validation of the tests employed, the hierarchical nature of the skills must be well established; otherwise the absence of a clear simplex pattern would be ambiguous: One could not tell whether the proposed hierarchy or the tests being used were at fault. It is also conceivable that a simplex pattern could be produced by a combination of faulty tests and non-hierarchically ordered skills. Thus, prior validation of the hierarchy is essential. Where compelling theoretical reasons to believe in hierarchical ordering exist [as, for example, in the Gagné (1968) learning hierarchy], the simplex technique could be used to validate tests of the skills in the hierarchy.

Where the development of the tests is undertaken with care such that the content of the skills is unambiguously represented in the tests, then validating the hierarchy should be possible using the simplex method. Guttman (1955) recognized this possibility and indicated the important new direction he felt this would provide for empirical investigation of learning phenomena:

> Previous experiments on learning seem to have emphasized largely the aspects of speed and difficulty. Simplex theory suggests study of certain aspects of the *organization* of learning. (p. 288)

However, certain caveats about the application of the simplex method, which Guttman has expressed, must be elaborated. The simplex pattern of correlations may be obtained when the ordering of tests is from least to most complex or vice-versa. Thus, no inference about the direction of the order of complexity is possible from this model. In instructional situations which derive from strong models of ordered complexity, such as the Gagné (1968) learning hierarchy, there will be other information about the direction of the ordering. The Taxonomy of Educational Objectives (Bloom, et. al., 1956) also provides an ordering of complexity which could serve to supplement the simplex method of analysis. As Guttman (1955) foresaw, however, in a less structured setting, the direction of ordering will not be certain. Furthermore, one cannot be sure whether any apparent ordering is inherent in the nature of the tasks or is an artifact of the learning process (Guttman, 1955). The discovery of a simplex pattern does not rule out the possibility of arranging instruction in a different way which would reverse the order of complexity. Only a very well developed learning theory, supported by carefully executed true experiments, can inform us whether this is possible or not. Guttman (1955) showed other ways in which components may combine to produce simplex patterns. Thus, the simplex is only one indicator, it is not a sufficient indicator of hierarchical ordering.

DISCRETE SCORE CASE

A model was derived by Murray (1971) for use in assessing developmental hierarchies. For this model, the score for each subject is summarized as a vector of pluses and minuses (minus for nonmastery of a level and plus for mastery of a level). The number of possible patterns of outcome is equal to 2^k, where k is the number of levels in the hierarchy. These 2^k patterns are listed and a frequency count of subjects having each pattern is obtained. For a four level hierarchy there are 16 patterns, listed in Table 1.

Those patterns marked with an asterisk are admissible under the hierarchical model. Clearly, the larger the proportion of subjects having these patterns, the more evidence there is to support the theoretical hierarchy.

Whereas, in the simplex model the linkages between skills were represented by shared variation (the off-diagonal elements of Σ_t), in this model the linkages are represented by conditional probabilities (p): The probability of scoring + on level i given that one scored + on level i - 1. Of course, p_1 is the unconditional probability of mastering the first level.

When the conditional probability, p_i, is equal to zero there is a break in the hierarchy: success on skill i is not dependent on success at level i - 1. When the conditional probability, p_i, is equal to 1, all subjects who succeed at skill i - 1 also succeed at skill i. The substantive implication is either that level i does not have a new component to acquire or that instruction at level i is perfect, guaranteeing that all who master level

Table 1. Patterns of Outcome for a Four Level Learning Hierarchy

Level			
A	**B**	**C**	**D**
*–	–	–	–
*+	–	–	–
–	+	–	–
–	–	+	–
–	–	–	+
*+	+	–	–
+	–	+	–
+	–	–	+
–	+	+	–
–	+	–	+
–	–	+	+
*+	+	+	–
+	–	+	+
+	+	–	+
–	+	+	+
*+	+	+	+

i – 1 will also master level i. There is also a possibility that the two adjacent tests or tasks have a special "linkage" due to their construction which operates independently of the substantive connection between skills. When p_i nears 1, the tasks involved and the learning situation must be carefully examined to determine what the likely cause might be.

Errors of measurement are represented in this model by a misclassification matrix, Q^i, for each skill $i = 1, \ldots, k$. For level i the matrix Q^i has the form:

$$(5)\ Q^i = \begin{bmatrix} Q^i_{11} & Q^i_{12} \\ Q^i_{21} & Q^i_{22} \end{bmatrix}$$

The rows of Q^i represent observed states and the columns represent true states. Thus, the probability that a person who has not mastered level i is classified as a non-master is Q^i_{11}. The parameter Q^i_{21} is the probability that a non-master will be classified as a master. These two parameters are constrained to sum to 1. Similarly, Q^i_{12} is the probability that a master is observed to be a non-master and Q^i_{22} is the probability that a master is observed to be a master. These two values must also sum to one. Thus, there are only two independent parameters in each matrix Q^i.

If the frequencies corresponding to the patterns of Table 1 are converted to a vector of proportions *P*, a model may be written in terms of the

Table 2. Parameterization of Latent Probabilities for the Outcome Patterns of 4 Level Learning Hierarchy

Level				Parameterization of π_n
A	B	C	D	
–	–	–	–	$\pi_1 = (1\text{-}p_1)$
+	–	–	–	$\pi_2 = p_1\ (1\text{-}p_2)$
–	+	–	–	$\pi_3 = 0$
–	–	+	–	$\pi_4 = 0$
–	–	–	+	$\pi_5 = 0$
+	+	–	–	$\pi_6 = P_1\ P_2\ (1\text{-}p_3)$
+	–	+	–	$\pi_7 = 0$
+	–	–	+	$\pi_8 = 0$
–	+	+	–	$\pi_9 = 0$
–	+	–	+	$\pi_{10} = 0$
–	–	+	+	$\pi_{11} = 0$
+	+	+	–	$\pi_{12} = p_1\ p_2\ p_3\ (1\text{-}p_4)$
+	–	+	+	$\pi_{13} = 0$
+	+	–	+	$\pi_{14} = 0$
–	+	+	+	$\pi_{15} = 0$
+	+	+	+	$\pi_{16} = p_1\ p_2\ p_3\ p_4$

misclassification parameters and latent probabilities of obtaining a particular pattern (π):

$$(6) \qquad P = Q\,\pi$$

where Q is the Kroneker product of the separate misclassification matrices:

$$(7) \qquad Q = Q^1 \otimes Q^2 \otimes Q^3 \otimes \ldots \otimes Q^k.$$

The structure of the vector π may be parameterized to represent the hierarchical model of interest. In the four level example under consideration there is a latent probability for each pattern of Table 1. These are to be parameterized in Table 2.

The parameters p_1, p_2, p_3, and p_4 are the conditional probabilities developed previously. The latent probability π of any unacceptable pattern under the hypothesized hierarchy is set to zero. The parameterization of the five acceptable patterns is explained below:

$\pi_1 = (1 - p_1)$:	The probability that a subject has mastered none of the levels of the hierarchy.
$\pi_2 = p_1(1 - p_2)$:	The probability that a subject has mastered level A but not mastered level B or beyond.
$\pi_6 = p_1\ p_2(1 - p_3)$:	The probability that a subject has mastered levels A and B but not C or beyond.

$\pi_{12} = p_1\, p_2\, p_3(1 - p_4)$: The probability that a subject has mastered levels A, B and C but not D.

$\pi_{16} = p_1\, p_2\, p_3\, p_4$: The probability that a subject has mastered all levels A through D.

The occurrence of unacceptable patterns is assumed to be due to misclassifications of acceptable patterns.

Murray (1971) showed how the preceding parameterization can be represented in the framework of the multinomial distribution and derives the maximum likelihood estimators of the parameters p_j and Q. The computational algorithm involves iterations on the first derivatives of the likelihood function with respect to the parameters. When these derivatives are acceptably close to zero, the parameter estimates may be displayed as well as an over-all test of goodness of fit and estimated standard errors of the parameter estimates.

Identification of the model requires the imposition of restrictions on the estimates of the misclassification probabilities. Not all matrices Q^j may have different parameter estimates.

The over all test of fit which contains N (the sample size) as a multiplier, is likely to be quite sensitive, especially if sample sizes of around 500 cases, as advocated by Murray (1971, p. 73), are adopted routinely. In this case the goodness of fit might best be assessed by comparing observed

Table 3. Observed Frequencies of Subjects by Pattern for Four Level Learning Hierarchy

Pattern	Level A	Level B	Level C	Level D	Number of Subjects with Pattern
*1	–	–	–	–	4
*2	+	–	–	–	20
3	–	+	–	–	1
4	–	–	+	–	0
5	–	–	–	+	1
*6	+	+	–	–	5
7	+	–	+	–	3
8	+	–	–	+	5
9	–	+	+	–	0
10	–	+	–	+	0
11	–	–	+	+	0
*12	+	+	+	–	17
13	+	–	+	+	2
14	+	+	–	+	4
15	–	+	+	+	0
*16	+	+	+	+	21

and reproduced probabilities for the outcome patterns. The model might be acceptable where the observed and reproduced proportions are nearly equal even though the likelihood ratio test would reject the model.

Clearly, one problem with this method is that the number of cells to observe and fit becomes large rapidly. A seven level hierarchy would have 128 cells. Further work needs to be done to determine whether arbitrarily dividing the hierarchy into smaller pieces would disturb the estimation of parameters and fit of the model.

EXAMPLE USING DISCRETE MODEL

An example of a four level hierarchy is worked out below using one of the data analyses in Sadek (1972). The hierarchy of skills was defined by the process due to Gagné, though in this analysis they are clustered into groups using a scheme alluded to by Gagné (1968). The four levels are:

Stimulus-Response → Discrimination → Rule Learning → Problem Solving

Table 3 shows the frequency of response for the 16 patterns as presented earlier. A high proportion of the cases (67 and 83) have acceptable patterns. For this analysis all four misclassification matrices were constrained to be equal. The parameter estimates are presented in Table 4.

The overall test of fit indicates that the proposed hierarchy fits the data quite well. We may turn our attention to the individual parameter estimates. The analysis indicates that a subject who has mastered the level is always correctly classified as a master. There is a 21% chance that a non-master will be classified as a master while 79% of the time the non-master is correctly classified as such. This may indicate that the criterion level of success for classification as master is set too low for these tasks. The probability of attaining mastery of the first level (stimulus response) is .91. The conditional probability of attaining mastery of the discrimination level given mastery of the stimulus response level is .52. The probability of mastering rule learning given mastery of discrimination is .91. This is quite high and may mean that rule learning, in this instance, is not distinct from discrimination. (Or it may mean that the instruction in rule learning given to subjects who master discrimination is nearly perfect.) In general, the decision as to whether this should be treated as a flaw in the measuring devices used or as an indication of unusually effective instruction could be determined by the curriculum and instruction experts managing the instructional program. Further evidence may need

Table 4. Parameter Estimates for Four Level Learning Hierarchy— Standard errors shown in parentheses

$\hat{p}_1 = .91\ (.04)$
$\hat{p}_2 = .52\ (.08)$
$\hat{p}_3 = .91\ (.10)$
$\hat{p}_4 = .46\ (.11)$

$$\hat{Q}' = \begin{bmatrix} .79\ (.03) & 0 \\ .21 & 1(.05) \end{bmatrix}$$

Goodness of fit: $\chi^2 = 7.15$ on 9 degrees of freedom

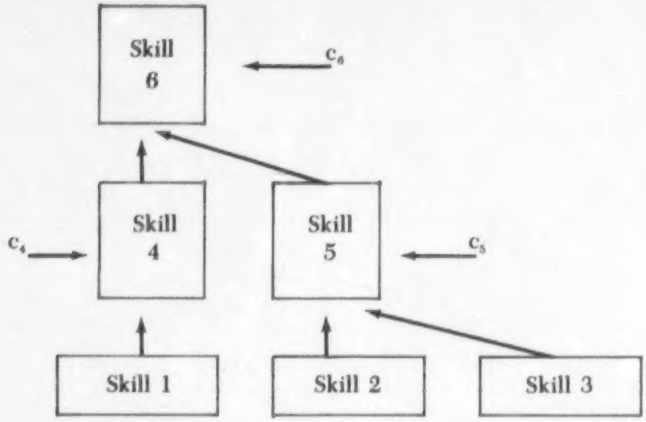


Figure 2: Complex of Ordered, Independent, Additive Components

to be collected. Finally, if rule learning is mastered, there is a probability of .46 of mastering problem solving in this context.

Goodman (1973) has developed a path analysis like model for discrete data which might be adapted to fitting hierarchies also. However, it is not explicitly parameterizable to account for misclassifications.

COMPLEX OF ORDERED, INDEPENDENT, ADDITIVE COMPONENTS

In this model, the relationships of skills at lower levels of complexity to those at higher levels are representable as hierarchically ordered "trees" but not as simple chains. The paper by Nitko in this monograph discusses examples of this nature. A simplified case is shown in Figure 2.

In words, figure 2 shows skill 4 to be composed of skill 1 plus a new component, c_4. Skill 5 is composed of two other skills (2 and 3) and a new component, c_5. Skill 6 is composed of skills 4 and 5 plus the new component, c_6.

CONTINUOUS VARIABLE CASE

We may write out a set of equations and specifications on the model as follows:

$$
\begin{aligned}
t_{j6} &= s_{j4} + s_{j5} + c_{j6} + e_{j6} \\
t_{j5} &= c_{j2} + c_{j3} + c_{j5} + e_{j5} \\
t_{j4} &= c_{j1} + c_{j4} + e_{j4} \\
t_{j3} &= c_{j3} + e_{j3} \\
t_{j2} &= c_{j2} + e_{j2} \\
t_{j1} &= c_{j1} + e_{j2}
\end{aligned}
\tag{8}
$$

where:

$$
\begin{aligned}
s_{j4} &= c_{j1} + c_{j4} \\
s_{j5} &= c_{j2} + c_{j3} + c_{j5}
\end{aligned}
$$

$$\begin{array}{ll}
\text{Covariance } (e_k\ e_m) = 0 & (k \neq m) \\
\text{Covariance } (e_k\ c_m) = 0 & (k, m + 1, \ldots, n) \\
\text{Covariance } (c_k\ c_m) = 0 & (k \neq m) \\
\text{Covariance } (e_k\ t_m) = 0 & (k \neq m)
\end{array}$$

Instead of writing out the covariance matrix for the observed scores, Σ_t, we will write, instead, a general matrix expression for the structure of this matrix. We begin with a representation of the i × 1 vector of scores on the measures for subject j, t_j:

$$t_j = (I - A)^{-1}\, c_j + e_j \tag{9}$$

Where I is the identity matrix, c_j is the vector of true scores on the components and e_j is a vector of errors. The content of the matrix A in expression (9) is detailed below.

The matrix A contains parameters representing links from one skill to another. These parameters are like regression coefficients in that they are asymmetric representations of the relationship between variables such that units of change of one variable are associated with units of change in the other. The correlation, by contrast, is unitless and symmetric—it does not change with the direction of the linkage. The matrix A corresponding to the model (8) is shown below:

$$A = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{array}
\begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0
\end{bmatrix} \end{array} \tag{10}$$

Here we have shown the parameters of A restricted to be equal to 1 because the statement of the model (8) shows this form. However, there is no reason to suppose that the regression of the true score part of one measure on another (e.g., c_4 on c_1) will be equal to 1. Furthermore, where the measures are in meaningful units (e.g., words per minute typed) a regression-type coefficient would be inherently interesting as well as having an important property of not varying with certain changes in population characteristics. Tukey (1954) has summarized the arguments in favor of these "structural" parameters. Therefore, a more general expression of (8) and (10) is shown below:

$$\begin{aligned}
t_{j6} &= a_4 s_{j4} + a_5 s_{j5} + c_{j6} + e_{j6} \\
t_{j5} &= a_2 c_{j2} + a_3 c_{j3} + c_{j5} + e_{j5} \\
t_{j4} &= a_1 c_{j1} + c_{j4} + e_{j4}
\end{aligned} \tag{8a}$$

$$t_{j3} = c_{j3} + e_{j3}$$
$$t_{j2} = c_{j2} + e_{j2}$$
$$t_{j1} = c_{j1} + e_{j1}$$

The same restrictions on covariances hold as in (8).

(10a)
$$A = \begin{array}{c|cccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ \hline t_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ t_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ t_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ t_4 & a_1 & 0 & 0 & 0 & 0 & 0 \\ t_5 & 0 & a_2 & a_3 & 0 & 0 & 0 \\ t_6 & 0 & 0 & 0 & a_4 & a_5 & 0 \end{array}$$

Solving for $(I-A)^{-1}$ shows how the true score part of t_j is derived from $(I-A)^{-1} c_j$:

(11)
$$(I-A)^{-1} = \begin{array}{c|cccccc} & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \\ \hline t_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ t_2 & 0 & 1 & 0 & 0 & 0 & 0 \\ t_3 & 0 & 0 & 1 & 0 & 0 & 0 \\ t_4 & a_1 & 0 & 0 & 1 & 0 & 0 \\ t_5 & 0 & a_2 & a_3 & 0 & 1 & 0 \\ t_6 & a_4a_1 & a_5a_2 & a_5a_3 & a_4 & a_5 & 1 \end{array}$$

Now we may write a general matrix equation for Σ_t:

(12) $$\Sigma_t = (I-A)^{-1} \Phi (I-A')^{-1} + \Psi$$

Where $(I-A)^{-1}$ is as developed above, Φ is a diagonal matrix of variances of the true score components, c_j, and Ψ is a diagonal matrix of variances of the errors of measurement, e_j. If the vector t is multivariately normally distributed with mean vector μ and covariance matrix Σ_t then maximum likelihood estimators of the parameter matrices of the model (12) may be developed following the lines investigated by Bock and Bargmann (1966) and Joreskog (1970). The estimation procedure, like that for the dichotomous response model of simple hierarchies in equation (6), consists of iterations on the derivatives of the likelihood function with respect to the parameters until these derivatives are acceptably close to zero. This provides a maximum likelihood solution for the parameters as well as estimated standard errors and an overall test of goodness of fit. A more detailed description of the procedure is found in Keesling (1972).

The reader will have realized that the simplex-generating hierarchy of Figure 1 is a special case of the model in (12). This means that the advantages of the technique just proposed (overall test of fit, estimates and standard errors of structural parameters) may be conferred on the simpler type of hierarchy discussed earlier. These parameter estimates should prove to be quite helpful in interpreting the outcome of the analysis. For example, when an estimate of an element of the matrix A, a_{k}, is zero then the hypothesized link between the measures is not substantiated by the data. Whether one decides to reexamine the hypothesized hierarchy or to review the construction of the measures will depend on which one is subject to the most doubt. The ratio of an element of Φ (a diagonal matrix) to the sum of that element and its counterpart in Ψ provides an estimate of the reliability of measurement of the component c_i unique to measure t_i. When a diagonal element of Φ is estimated to be zero, then the inference is that there is no unique component c_i measured by t_i. Here, the nature of the measure, the hypothesized orderings, and the instructional setting all play a part in the interpretation. The statistical model cannot distinguish among these sources as causes of the problem.

If the measures t_i were sufficiently well constructed as to have no measurement error it would be possible to use ordinary least squares to estimate the parameters of A and Φ in (12). Further work needs to be done to determine whether there is a point at which error variance is small enough to be considered zero in the context of these models.

The problem of identifying the parameters in the model is discussed at length by Keesling (1972). By identification we mean that we must have appropriate information to estimate the parameters. In the case of the simplex-type hierarchy we would have to restrict the first and last elements of Ψ to be equal in order to separate the error variances from the true score variances. There may be circumstances where this restriction will lead to the failure of the model to fit the data. In other circumstances, however, there may be reason to restrict all Ψ's to be equal in the estimation. Indeed, one might wish to restrict certain parameter estimates to particular values depending upon the nature of the measuring instruments and information obtained in previous studies with the complex of skills. In general, the identification of parameters may be made with fewer constraints when there are replicate measures for each skill in the complex. Keesling (1972) shows an example of replicate measures in the context of an analysis of school and community factors in achievement.

Certain difficulties may arise where some skills at the lowest level are apparently acquired by all subjects. In this instance the associated variance parameter Φ_{ii} will be estimated to be zero and links from this skill to others, represented in A, will be unidentified. The analysis will have to recommence, eliminating these skills. Further evidence will have to be obtained to determine whether this is an effect of the instruction or the measures used.

When the sample size is large it may happen that the overall test of goodness of fit will indicate a statistically significant lack of fit where the observed covariance matrix of *t*, S_t, is reasonably well reproduced

Table 5. Acceptable Patterns and Parameterizations for Model of Figure 2

Pattern							Parameterization of Latent Probability, π
Skill #	1	2	3	4	5	6	
	−	−	−	−	−	−	$\pi_1 = (1-p_1)(1-p_2)(1-p_3)$
	+	−	−	−	−	−	$\pi_2 = p_1(1-p_4)(1-p_2)(1-p_3)$
	−	+	−	−	−	−	$\pi_3 = p_2(1-p_3)(1-p_1)$
	−	−	+	−	−	−	$\pi_4 = p_3(1-p_2)(1-p_1)$
	+	+	−	−	−	−	$\pi_5 = p_1(1-p_4)p_2(1-p_3)$
	+	−	+	−	−	−	$\pi_6 = p_1(1-p_4)p_3(1-p_2)$
	−	+	+	−	−	−	$\pi_7 = p_2p_3(1-p_5)(1-p_1)$
	+	+	+	−	−	−	$\pi_8 = p_1p_2p_3(1-p_4)(1-p_5)$
	+	−	−	+	−	−	$\pi_9 = p_1p_4(1-p_2)(1-p_3)$
	+	+	−	+	−	−	$\pi_{10} = p_1p_4p_2(1-p_3)$
	+	−	+	+	−	−	$\pi_{11} = p_1p_4p_3(1-p_2)$
	+	+	+	+	−	−	$\pi_{12} = p_1p_4p_2p_3(1-p_5)$
	−	+	+	−	+	−	$\pi_{13} = p_2p_3p_5(1-p_1)$
	+	+	+	−	+	−	$\pi_{14} = p_1p_2p_3p_5(1-p_4)$
	+	+	+	+	+	−	$\pi_{15} = p_1p_2p_3p_4p_5(1-p_6)$
	+	+	+	+	+	+	$\pi_{16} = p_1p_2p_3p_4p_5p_6$

by the estimated parameters of the model in (12). Therefore, when the test indicates lack of fit one should attend to the discrepancies between the observed and reproduced covariance matrices as a way to judge the importance of the lack of fit. A variant of the statistic suggested by Tucker and Lewis (1973) could also be used to assess the acceptability of the model.

DICHOTOMOUS DATA

A structure like that of Figure 2 representing six skills would produce a table like Table I with 64 possible outcomes rather than 16. Of these 16 would be acceptable latent patterns, given the model of Figure 2. These patterns and the parameterization of the latent probabilities associated with each are shown in Table 5. The method outlined in Murray (1971) would be used to obtain estimates of the probabilities, p_i.

THE COMPLEX OF ORDERED, INTERRELATED, ADDITIVE COMPONENTS

Here, the basic model is very similar to that presented in Figure 2 and equations (8). However, we now remove the constraint that the covariances of the c_i be equal to zero. Thus, c_4 and c_5 of that model may be correlated, for example. Skills 4 and 5 may be similar in some respect even though the prerequisite skills are dissimilar.

The analytic model for the continuous case (12) may be extended to cover this case by allowing Φ to be a general, symmetric, positive definite matrix (rather than a diagonal matrix). The method of estimation given in Keesling (1972) will estimate the parameters of this model. However, identifying parameters in this circumstance can be quite tricky. Unfortunately, there are no easy ways to determine which parameters of a model are underidentified. Keesling (1972) presents two usable, but tedious, methods.

Murray's model may be expanded to allow for associations among components by including joint probabilities in the parameterization of the model. In both analytic frameworks it is very important to consider the substantive implications of such associations. For example the inclusion of an association between c_4 and c_5 of the model in (8) would seem to imply the existence of an unmeasured component influencing both. Presumably, the theory of instruction would be strong enough to rule this out a priori, otherwise this component ought to be built into the model explicitly and the performance of subjects on this component should be assessed.

A further possibility would be that data could be collected in the form of a time series of measurements on each skill. A very general sketch of models which can be applied to longitudinal data is given by Murray, Wiley and Wolfe (1971). Some variants of longitudinal factor analysis models might also be appropriate. (See Corballis & Traub, 1970, for example.) Murray (1971) shows models for some longitudinal data taken from Anderson (1955) as well as longitudinal data arising from research on Piaget's developmental hierarchies.

CONCLUSION

The effects of a faulty instrumentation of criterion-referenced measures may be evident in several forms. Where the objectives to be learned are not connected by a priori orderings there are rather straightforward tests of independence available. When, in addition, information is available about the performance of the subjects on a "criterion," then there are clear methods of assessing the validity of the tests as predictor of that criterion performance. Special notice should be taken of instructional history in which the "criterion" may be manipulated and used to predict performance on the criterion-referenced test. Ozenne (1971) has made use of this notion.

In the context of the related objectives models, we note that the validity of the proposed structure of relationships among objectives is very important. Where our theory of instruction is strong enough for us to accept a proposed a priori ordering of objectives, we may use the methods suggested earlier to assess the validity of the criterion-referenced measures. A faulty measure may show up as an absent link between objectives; one or both measures did not assess the common components the theory of instruction attributed to each objective. A faulty measure may be evidenced by an apparent absence of one of the hypothesized components, c_i. In the continuous case this is a zero estimate for a diagonal element of Φ while, for the dichotomous case it is an estimated p equal to one. Either the measure used to assess the higher level skill does not contain the new component or the measure(s) at the lower level does contain this component. Clearly, the domain of each objective must be well specified in order to avoid this type of difficulty.

Unfortunately, there are other difficulties which may produce similar parameter estimates or may be reflected in a general lack of fit at a specific point or overall. These include forgetting and unforseen paths to higher level skills.

Forgetting occurs where a subject has learned the levels in the hypothesized order but has forgotten how to perform lower level tasks while retaining higher level tasks. (As White, 1973, pointed out, if he learns the higher level skill first and then forgets it while retaining the lower level skill, the evidence he contributes will favor the hypothesized hierarchy. White suggested testing acquisition of skills continuously throughout instruction as a way to control for this occurrence.) If lower level skills should still be part of the repertoire of subjects who show mastery of higher levels, as Murray (1971) presumed in his presentation of developmental hierarchies, then evidence of forgetting may seriously discredit the learning hierarchy notion.

The statistical models we have proposed assume that the hypothesized ordering is the sole route to acquisition of higher order skills. However Gagné (1968) stated:

> A learning hierarchy, then, in the present state of our knowledge, cannot represent a unique or most efficient route for any given learner. Instead, what it represents is the most probable expectation of greatest positive transfer for an entire sample of learners concerning whom we know nothing more than what specifically relevant skills they start with. (p. 3)

He further identified "skipping" of intermediate levels, transfer from another learning domain, and "atypical combinations of subordinate skills" which for some learners produce learning transfer as possible alternate routes to higher level skills. All of these occurrences would yield the same result in the context of the statistical models proposed above; data which do not fit the model.

Thus, it will fall to the data analyst—who must also understand the development of the learning hierarchy at issue, the tasks used to represent its levels, and the instructional history of the subjects who are used in the validation study—to assess the fit of the statistical model to the data and to make a "good guess" about the nature of any observed lack of fit.

REFERENCES

Anderson, T.W. Probability models for analyzing time changes in attitudes. In P.F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, Ill.: Free Press, 1955.

Bloom, B.S. (Ed.) *Taxonomy of educational objectives, handbook I: Cognitive domain.* New York: David McKay, 1956.

Bock, R.D., & Bargmann, R.E. Analysis of covariance structures. *Psychometrika,* 1966, *31,* 507-534.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81-105.

Corballis, M.C., & Traub, R.E. Longitudinal factor analysis. *Psychometrika,* 1970, *35,* 79-98.

Gagné, R.M. Learning hierarchies. *Educational Psychologist,* 1968, *6,* 1-9.

Goodman, L.A. The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika,* 1973, *60,* 179-192.

Goodman, L.A., & Kruskal, W.H. Measures of association for cross classifications, IV: Simplification of asymptotic variances. *Journal of the American Statistical Association,* 1972, *67,* 415-421.

Guttman, L. A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, Ill.: Free Press, 1955.

Joreskog, K.G. A general method for analysis of covariance structures. *Biometrika,* 1970, *57,* 239-251.

Kaiser, H.F. Scaling a simplex. *Psychometrika,* 1962, 27, 155-162.

Keesling, J.W. Maximum likelihood approaches to causal flow analysis. Unpublished doctoral dissertation, University of Chicago, 1972.

Morrison, D.F. *Multivariate statistical methods.* New York: McGraw-Hill, 1967.

Mukherjee, B.N. Derivation of likelihood-ratio tests for Guttman quasi-simplex covariance structures. *Psychometrika,* 1966, *31,* 97-124.

Murray, J.R. Statistical models for qualitative data with classification errors. Unpublished doctoral dissertation, University of Chicago, 1971.

Murray, J.R., Wiley, D.E., & Wolfe, R.G. New statistical techniques for evaluating longitudinal models. *Human Development,* 1971, *14,* 142-148.

Ozenne, D.G. Toward an evaluative methodology for criterion-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1972.

Sadek, C.S. A study of transfer in foreign language learning. Unpublished doctoral dissertation, University of California, Los Angeles, 1972.

Schönemann, P.H. Fitting a simplex symmetrically. *Psychometrika*, 1970, *35*, 1-21.

Tucker, L.R., & Lewis, C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 1973, *38*, 1-10.

Tukey, J.W. Causation, regression and path analysis. In O. Kempthorne, et al., (Eds.), *Statistics and mathematics in biology*. Ames, Iowa: State College Press, 1954.

White, R.T. Research into learning hierarchies. *Review of Educational Research*, 1973, *43*, 361-375.